Guidance for Evaluating Mass Communication Health Initiatives: Summary of an Expert Panel Discussion

Held May 3-4, 2004
Atlanta, GA

Sponsored by:
The Centers for Disease Control and Prevention
Office of Communication

Edited by:

May G. Kennedy, Ph.D., M.P.H.
CDC Division of Health Communication

Jodie Abbatangelo, M.A., Sc.M.
CDC/ORISE Fellow

May, 2005

The authors would like to express appreciation to Dr. Galen Cole and Mr. Tom Chapel for helping to plan this discussion, Ms. Vivian Jones for handling participant travel and logistics, Dr. Christine Prue for facilitating the electronic focus group discussion, Dr. David Cotton for facilitating the oral discussion, and especially to the outside expert panelists (*listed in appendix A*) for their thoughtful contributions.

Table of Contents

Background

The burden of disease in the United States can be lessened if members of the public practice certain behaviors. Hundreds of millions of dollars have been spent in the last few years alone on large-scale health communication campaigns and community-wide health communication programs designed to promote health-related behaviors. The creative formats and new electronic media employed in many of these communication programs have intuitive appeal, but evidence of program effectiveness is needed to inform future practice and to track the public health return on a major investment.

Evaluating large-scale communication programs presents a host of challenges (Valente, 2002). On May 3-4, 2004, CDC's Division of Health Communication (DHC) convened a panel of experts to clarify and suggest ways to approach some of the relevant issues. This report summarizes the discussion.

The experts (*see Appendix A*) came from private, academic, and governmental sectors, and represented a wide range of substantive areas. Their advice will be used by DHC to improve products and activities such as (1) CDCynergy, a CD ROM-based interactive tool for planning, implementing and evaluating communication programs, (2) evaluation course curricula, and (3) one-on-one technical assistance to CDC staff and external prevention partners.

An electronic focus group facility was used to maximize each panelist's opportunity for input, but the facility could accommodate only a small number of panelists. Consensus on many of the issues will require broader discussion and additional scientific progress. This report is intended to articulate and explore some of the key issues in health communication program evaluation and to encourage others to join in the dialogue.

Consultation Procedure

Experts scheduled to participate were sent possible points for discussion and asked to rate their importance. Qualitative comments were also requested. In judging the importance of a topic, participants were instructed to consider the breadth of its applicability to health communication evaluation, recent innovations in the area, and the feasibility of incorporating pertinent advice. Ratings and comments were also solicited from CDC staff members who have conducted evaluations of health communication programs.

The two sets of feedback were compiled and found to be very similar, and the discussion guide was modified accordingly. General topics retained included (1) the role of theory in the evaluation of health communication programs and campaigns, (2) recommended evaluation methods, strategies and statistical issues, (3) coping with budget constraints, (4) considerations pertaining to health disparities and minority populations, and (5) public policy. A summary of the expert panel discussion is organized under these topics in the following section of this report.

Before the panel was assembled, several documents (*see Appendix B*) were sent to participants to provide a common frame of reference for the discussion. Attention was drawn to a report based on a USAID-funded conference of evaluators of communication programs outside the United States, especially in developing countries (Figueroa, Bertrand, & Kincaid, 2002). Commonly referred to as "The Belmont Report," it provided a point of departure for examining evaluation in health communication circumstances that do not lend themselves to "gold standard" randomized designs.

On the mornings of May 3rd and 4th, participants met at CDC in Atlanta in an electronic focus group facility. It consisted of 12 computer terminals for participants and a moderator's station, all in one room. Group Systems Software allowed the moderator to control the length of time that each question was displayed on a large screen. All participant responses appeared on the large screen in real time and were retained in electronic files. Transcripts were printed immediately after the focus group sessions.

CDC staff joined panel members for verbal discussions each afternoon. Transcripts of the morning's online session were distributed, and a moderator recapped the key points raised, sought clarifications, and encouraged panelists to elaborate on their previous written input. At the end of the afternoon session on the first day, CDC staff and the external experts formed three breakout groups to exchange perspectives on factors that affect the practice of health communication practice within the federal system. At the end of the afternoon session on the second day, the experts made evaluation policy recommendations and suggested next steps in eliciting expert advice.

The transcripts from the morning sessions and the notes on the afternoon discussions were summarized into a draft report. The draft was distributed to the expert panelists for corrections and comments, and their feedback was incorporated into the present version of this report. All the contributors hope that it will be useful to evaluation practitioners.

Expert Panel Discussion[1]

I. THEORY

The expert panelists offered their views about (1) when evaluators should use formal theory and when they should use a logic model, (2) the contribution a theorist makes to an evaluation effort, (3) underutilized theories, (4) the advisability of using one intact theory vs. components of several theories, and (5) what to do when outcome data suggest a breakdown in the causal pathway. This section of the report provides the gist of their comments about theory, along with references to literature that offers more detail.

Theories and Logic Models

Behavior change theory (Glantz, Lewis & Rimer, 1997; NCI, 2003) explains how psychosocial determinants of health interact to spur behavior change, without reference to

[1] When a theme is presented without comment about the extent of panel member agreement on the matter, general agreement should be assumed. Minority opinions and disagreements are noted.

a communication campaign. There is current confusion within the communication field between this kind of formal theory and a "theory of the project" or logic model (Yin, 2003; Cole, 1999; Chen, 1990; Shadish, Cook & Leviton, 1990). Normally presented as a flowchart, the logic model outlines hypothesized causal pathways through which effects are expected to come about. It includes not only constructs from one or more theories but also program elements that are expected to bring about change. One expert argued that the term "theory" sufficed to cover this kind of model, but much of the discussion preserved the distinction between theories and logic models.

The logic model should be detailed enough to guide decisions about which data are gathered, from whom, and how often. It should specify not only the elements of the intervention and its desired outcomes, but also intermediate variables in the causal chain and external factors that could be rival explanations for any behavioral change observed. An initial version of the evaluation plan should be broadly outlined on the basis of the logic model. The level of information that decision-makers actually will use and the size of the evaluation budget should be considered in refining the plan (see Andreasen, 1985, on 'backwards' research).

A relatively long term, big budget project often requires an elaborate, dynamic logic model that will evolve with the project. Specifying the theory of the project is an iterative process that should begin at the earliest stage of campaign planning and continue as new information becomes available. Armed with the preliminary outline of a model based on behavior change theory, epidemiology, and experience in similar programs, the evaluator can begin to identify target audiences. The audience perspective gleaned through subsequent formative research will, in turn, inform revisions in the logic model, and thus in the evaluation design.

The more detailed the logic model is, the more useful it will be to evaluators and program planners, but the greater the chance of updating adjustments will be. Logic models can be refined right up until evaluation instruments are in final form and scheduled for immediate use, and some experts suggested making ongoing adjustments even after a campaign is launched.

Because most health communication campaigns and community-wide programs are conducted to improve health, not to test theory, the logic model should drive evaluation design. At the same time, evaluation results can spur theoretical progress when the logic model incorporates a sufficient number of elements from one or more formal theories.

<u>A Theorist's Role in Evaluation</u>

The evaluation research team typically consists of program staff in charge of program planning and a program evaluator. The program evaluator is often the only behavioral or social scientist on a communication project. Unless the evaluator wears the "hat" of theorist too, the theory of the project is likely to remain implicit at project launch. The failure to make the logic model explicit risks wasting resources on message strategies that

are not adequately linked to psychosocial predictors of behavior, and on performance measures that are off mark.

Whether or not program personnel are trained in theory, they should be involved in logic model development so that the theory of the project is grounded in more than evaluator assumptions. Such teamwork can be difficult -- evaluators and program planners often speak different languages -- so evaluators should learn to conduct discussions about how the program will work in non-theoretical terms. Asking program planners for a general description of program outcomes of interest is a good way to begin to develop a logic model. Productive dialogue is fostered by an up-front understanding that there will be a series of logic models in which versions become increasingly more specific and complex as theoretical frameworks are "tried on for size" and theory-driven objectives are defined.

As the primary author of the logic model, the evaluator plays a key role in guiding expectations about program performance. Showing funders how campaign activities operationalize theory can contribute not only to securing adequate resources to conduct the program and its evaluation, but also to "right-sizing" outcome expectations. Some experts recommended that evaluators craft two versions of the logic model. A relatively simple version could be available for policy-makers who are interested mainly in behavior change objectives. It should specify the program time frame and how each change leads to the next level of change. Another version could provide the detail that operations people need, including how change will be detected and the steps to take if the desired change is not achieved at the desired level on the projected timeline.

In instances where theory specification was weak initially and the evaluator is introduced to the project after the design or implementation of the intervention, it is appropriate for the evaluator to work with program planners to fill in the gaps post hoc. In motivating staff to flesh out the logic model, the evaluator should point out that sufficient detail about the processes of change will make it possible to craft good performance measures and explain them to stakeholders.

Underutilized and Underdeveloped Theories

Long-running programs should adopt research agendas in which logic models with strong theoretical bases are developed and tested over time. Although more basic theoretical work is needed, communication program evaluators do not take full advantage of a rich existing body of potentially germane theory (DiClemente, Crosby & Kegler, 2002). This is due, in part, to the reality that most evaluations are conducted to document the effectiveness of programs, not to understand how they work. Evaluators should assess the pertinence to their work of the theoretical approaches listed below.

Individual-level theories of behavior change such as the Theory of Reasoned Action (Ajzen & Fishbein, 1980), the Transtheoretical Model (Prochascha, Diclemente & Norcross, 1992), and Social Cognitive Theory (Bandura, 1986) are used frequently in the evaluation of health communication programs, but some types of individual-level theories have been neglected. Too little attention has been paid to theories about how

genetic factors determine an individual's reactions to health messages and other environmental stimuli (e.g., Palmgreen, Donohew, Lorch, Hoyle & Stephenson, 2001). More focus on theories of information processing (e.g., McGuire, 1981; Williams-Piehota, Schneider, Pizarro, Mowad, & Salovey, 2003; Rothman & Salovey, 1997) is warranted. There should be special emphasis on communication theories that address attention, attitudes and subjective norms such as the Elaboration Likelihood Model (Petty & Cacioppo, 1979), attitude accessibility theory (e.g., Fazio & Williams, 1986), and attitude-to-the-ad theory (Shimp, 1981). Reactance theory (Brehm, 1966) and others that examine the effects of emotion on communication have been underutilized. The ideation model (Kincaid, 2000a, 2000b; Storey, Boulay, Karki, Heckert & Karmacharya, 1999) that has been used primarily overseas in the evaluation of family planning communication programs takes cognitive, emotional and social factors into account. An inclusive model of this kind may be useful in guiding evaluation of communication programs in the United States.

Evaluators here often limit their focus to one-way, multi-step "flow" models, and tend to ignore mediators and moderators of behavior. As a result, message exposure, reception and interpretation are poorly understood and often assumed.

In fact, with the possible exception of the theory of diffusion of innovation (Rogers, 1995), evaluations of domestic communication efforts have failed to take advantage of theories that go beyond the level of individual behavior change (see Glantz, et al., 1997, for several examples of supra-individual level theories). These theories explicate the effects of cultural factors (e.g., technology) and language, as well as mapping interactions among target audience members, intermediaries (e.g., sexual partners, medical service providers), and social structures (e.g., families, peer groups, class, religion, public policies). Because such interactions are the context of individual message reception, they should be brought to bear in conceptualizing campaign effects on individual audience members. Moreover, certain intervention strategies (e.g., media advocacy) may change the contextual factors.

Panelists suggested that social network theory (e.g., Kawachi & Berkman, 2001) and network analysis (Kincaid, 2004) can sometimes illuminate the social dynamics within which communication operates. They also suggested that problematic integration theory (PIT) (Babrow, 2001) could offer insight into selective exposure by drawing attention to the costs and benefits of exposure.

Whole categories of theory that hold promise for evaluators of communication projects were also highlighted by panel experts. They included theories about: evaluation itself (Stufflebeam, 2001; Cole, 1999; Shadish, et al.,1990; Chen, 1990), public relations (e.g., Grunig & Hunt, 1984), social interaction (e.g., Grunig & Turner, 1989), political influences, participation, social action and social change (e.g., Foster-Cohen, 2004; Himmelstrand, 1981; Blum, 1976; Kondo, 1975), fear management (Witte, 1992), social processes of change (e.g., Kincaid, 2004; Anderson, 2000; Terry & Hogg, 1999), the cultivation of perceptions of reality through television viewing (e.g., Gerbner & Gross, 1976), chaos (e.g., Murphy, 1996), concepts of control over events or optimistic bias

(e.g., Wein, 1987), leadership development (House & Aditya, 1997), and differential effects of interventions on specific subgroups (e.g., Martin, 2004; Dearing, 2004; Ridgeway & Correll, 2004; Sun, 2003; Carli & Eagly, 2001; Plewczynski, 1998; David & Turner, 1996).

Finally, very broad, multi-level theories such as systems theory (e.g., Von Bertalanffy, 1967) and the social ecological framework (Bronfrenbrenner, 1979; Green & Kreuter, 1991) were identified as useful structures within which various theories can be embedded. A broad evaluation framework can take into account more narrowly focused theories, either in whole or in part, and can accommodate secular change and other factors that have the potential to influence intervention effects.

There were two mentions of areas in which more basic theoretical work is needed. One panelist expressed the opinion that language comprehension theories are insufficiently well developed to guide media literacy approaches. Another panelist felt that models of interactions among multiple systems levels require conceptual work before they can be usefully applied to the evaluation of heath communication campaigns and programs.

Keeping a Theory Intact

Questionnaires intended for target audiences stand out as the "work horses" in the field of measuring communication program effectiveness. Understandably, questionnaire respondents have limited time, attention and tolerance for answering questions, so space on evaluation instruments is at a premium. This space shortage creates a trade-off in which guarding the integrity of any one theory (by measuring all of its component constructs) is pitted against the inclusion of constructs from other theories that provide alternate explanations of behavior change. The dilemma is compounded by the fact that multi-item scales are often necessary to measure theoretical constructs reliably.

Panelists assigned different weights to the advantages of measuring one theory fully, but there was consensus that employing only intact theories was neither necessary nor advisable. However, they agreed that a theory must be measured very well if a project is to be designed on that single theory alone. Costs and benefits of using a single theory are listed in Box 1.

Box 1

Advantages of keeping a theory intact:

1. Measuring all the constructs along the causal pathway specified by a theory allows the evaluator to (a) explain how a program worked (when it worked), (b) to point to promising intermediate outcomes (if they were observed) when behavioral outcomes fell short of objectives, and (c) to diagnose areas in which the intervention was weak if there was no observed change in measures of the intermediate constructs.
2. Data about theory performance in predicting outcomes can be used to improve the theory.
3. Use of a single theory in both message design and evaluation gives the project conceptual coherence.
4. A logic model framed in terms of a single theory is parsimonious and easier to explain to decision-makers.
5. Some theories have outperformed others so it is efficient to use them.
6. Some behavior changes require relatively simple interventions for which a single theory is adequate.
7. Competing theories can imply mutually exclusive evaluation design criteria. For example, differing predictions about lag times between exposure and outcomes would dictate different schedules for survey waves.

Advantages of a multi-theory approach:

1. Single theories tend to specify relationships at one systems level, and change is influenced and observed at multiple levels. Complicated interactions are best represented by components of multiple theories under a broad systems theory framework.
2. Mediators can vary by audience subgroup. For example, the racial discrimination experienced by some subgroups may mediate behavioral outcomes and may not be represented by a basic theory of behavior change.
3. Using more than one theory is appropriate when there are equally viable alternate hypotheses for behavior change prior to the intervention.
4. Behavior change can be observed when there was no apparent change in theoretically antecedent constructs, or when there was a break along the anticipated causal chain. In these circumstances, having measures of constructs from competing theories may illuminate the processes of change.
5. Theories often have overlapping constructs, allowing different causal pathways to be represented in minimal space.
6. When the task is to evaluate a program, not to test theory, an ecumenical approach that picks and chooses constructs from various sources allows maximum flexibility to adapt measures to a particular context.

To help evaluators resolve the space dilemma in the future, behavioral research leading to the development of short forms of scales of important constructs was recommended. One panelist also advocated for research on how theories complement each other.

Breakdowns in the Causal Pathway

When evaluation data are analyzed, it is more the rule than the exception to find a breakdown somewhere along the theorized causal chain. Hypotheses about how the

program would work may have been incorrect, a current event or secular trend[2] may have overwhelmed program effects, measures of intermediate constructs may have been insensitive, or analytic techniques (e.g., failing to control for multicollinearity, the overlap between predictor variables) may have obscured real effects.

Although programs rarely function exactly as predicted, evaluators should exercise particular caution in their interpretations when the broken link in the causal chain is between message awareness and theorized outcome antecedents such as attitudes, beliefs and intentions. Evidence already in the literature may help the evaluator determine whether there was a bad fit between reality and the proposed theoretical pathway or the constructs were poorly operationalized. Depending on what was measured, it may be possible to explore the power of alternative theoretical explanations statistically. In some instances, media monitoring could reveal that an event outside of the program actually led to the outcome (e.g., the announcement that a president's wife has breast cancer led to an uptake of mammogram services).

If the program achieves its behavior change objectives, one is tempted to celebrate success and not worry too much about a lack of empirical support for the initial mechanism of change posited. However, a lack of understanding of how a change happened means that the change is unlikely to be replicated.

II. METHODS

Evaluation methodology was the second major area addressed by the expert panelists. This part of the discussion began with a question about the appropriateness of various study designs for evaluating large-scale communication programs. The panelists were then asked for comments on a series of specific design challenges ranging from measuring secular trends to increasing the reliability of self-reported campaign exposure. The panelists made a number of general comments about statistical analysis, and gave advice on strategies for imputing missing data.

Methods: Design Issues

The Infeasibility of Randomized Designs

The true experimental design (or randomized controlled trial [RCT] as it is called in the medical and epidemiological literature), is said to provide the best evidence for public health intervention efficacy because of its ability to attribute cause for observed effects to the intervention (Zaza, et al., 2000). However, random assignment is impossible in evaluations of national communication campaigns and can be impractical or unethical in designs for community-wide programs (Valente, 2002).

[2]Trends in health indicators and their determinants that emerge over time, irrespective of the effects of any single programmatic intervention.

Barriers to using RCTs as tests of "full coverage" programs with media components include:

- the prohibitive cost and analysis challenges[3] of randomizing at the community level of analysis,
- the difficulty of limiting exposure to a specific target audience in a community,
- community resistance to no-treatment control status when an intervention has face validity and the health issue is urgent,
- activities in the control community that parallel intervention activities, and
- a common lack of generalizability of results to other communities.

When a campaign cannot be evaluated with an RCT, it may be possible to use experimental procedures to pilot-test the efficacy of the planned communication strategy in a small geographic area prior to the formal campaign launch. Efficacy pilots are ethically and fiscally responsible, but have been omitted in the past because of political pressure to respond rapidly and at full scale to health issues of widespread concern. Several panelists argued that pilot-testing programs before they are rolled out is a conscientious use of resources, but one expert doubted that the results of a small-scale RCT would alter a campaign strategy that had been predetermined by policy makers.

When efficacy testing is not possible, there are ways to introduce an element of randomization into the evaluation scheme. For example, exposure could be forced among a small subsample in a controlled, experimental fashion. One panelist suggested systematically varying dosage across media markets over time such that communities initially in a control condition would, by the end of the study, receive an "equitable" dose. Another expert voiced skepticism that media could be manipulated in this way.

Non-randomized Designs

A number of quantitative but non-randomized designs can be appropriate for evaluating large communication campaigns. They include cross-sectional, cohort, time-series and quasi-experimental designs.

When randomization is infeasible even though a program is not national in scope, it may be possible to identify a community that is similar to the intervention community and to conduct a quasi-experiment, preserving what one expert called "some claim" to internal validity. Of course, quasi-experiments face many of the same barriers that confront RCTs. Both the control sites in RCTs and the comparison sites in quasi-experiments can be contaminated by direct or indirect exposure to campaign messages. After the campaign is already underway, historical events (e.g., a disease outbreak) can render a comparison or control site non-comparable to an intervention site. Assessing the quality of program implementation carefully and collecting other process data can help to rule out these possibilities. Utilizing cross-over designs in quasi-experiments can help to address validity problems and ethical concerns about withholding a valued intervention

[3]Adjustments for group clustering may be made by using Hierarchical Linear Modeling (HLM) software; see Murray (1998) for prescriptions regarding degrees of freedom and effects that must be tested.

from a comparison community (Palmgreen, Donohew, Lorch, Hoyle & Stephenson, 2002).

Cross-sectional surveys are a frequently chosen evaluation design, in part because several studies that have employed both cross-sectional and cohort surveys have found stronger program effects in the cross-sectional data (Snyder & Hamilton, 1999). However, there are two major threats to the internal validity of cross-sectional surveys: the potential for self-selection bias with regard to campaign exposure[4] and endogeneity.[5]

An evaluator can apply statistical controls defined in the following section for these threats to validity, and should reinforce and supplement cross-sectional designs with rigorous formative evaluation and careful qualitative process evaluation. After the campaign, historical and ethnographic data collection methods can provide information to confirm the results of cross-sectional surveys. For example, one could return to the field and ask members of the target audience whether program evaluation results were an accurate representation of their lived experiences. The low-cost evaluation strategies outlined later in this report also hold promise for confirming evaluation results from cross-sectional surveys.

Even when statistical controls are applied and confirmatory data collection is undertaken, some journal editors and scientifically sophisticated policy makers regard anything other than an RCT to be a weak substitute. The terms used for non-randomized designs may contribute to their devaluation. For example, calling them "quasi-experimental" or "alternative" implies that these methods are quasi-scientific. Although they do not permit the same degree of confidence in causal inferences that RCTs do, quasi-experiments are not quasi-scientific. Developing criteria for the conduct and description of designs that are more compatible with mass communication research may help them become more acceptable to prestigious journals and, in turn, to some decision-makers (see Des Jarlais, et al., 2004, for a suggested description scheme).

The assumption that RCT evidence is the only valuable kind can be changed gradually by accumulating evidence from a variety of sources to support the findings of non-RCT studies. Consistent findings across studies that used multiple methods with complementary strengths can be compelling.

Previous Advice on Designing International Campaigns

In order to develop guidelines for evaluators of communication programs that the United States Agency for International Development funds overseas, the agency convened a group of experts in 2001 at the Belmont Conference Center in Elkridge, Maryland

[4] Self selection bias occurs when exposure to messages in a communication program is not random across a population. Instead, exposure is a function of pre-existing characteristics of individuals; these characteristics may be associated with desired program outcomes.

[5] In general, endogeneity refers to reciprocal causation of variables within a model. Here, as in the Belmont report (see page 6), it refers to uncertainty regarding the causal direction between exposure to the program and the observed outcome.

(Figueroa, Bertrand & Kincaid, 2002). The experts made eight recommendations for reinforcing causal claims when non-RTC designs are used (*see Appendix C*). Randomization ensures that true experiments meet the first four criteria but not the last four.

In 2004, the CDC expert panelists were asked whether the recommendations in the Belmont report are necessary and sufficient for eliminating threats to the internal validity of evaluation claims made about full coverage programs in the United States. Some experts felt that the criteria were necessary in most cases but, in the aggregate, probably insufficient.[6] Others felt that they were consistent with epidemiological standards for causal inference (see Hennekens & Buring, 1887) and considered them sufficient if all met. There was consensus that they should be viewed as design features to consider when building an evaluation to meet the unique needs of a particular campaign.

The Belmont report recommended using cross-sectional survey designs along with statistical methods such as propensity scores[7] (Rosenbaum & Rubin, 1983; Bollen, 1995) to control for self-selection and endogeneity. One member of the expert panel convened by CDC agreed that propensity scoring was a "practical and statistically equivalent" alternative to RCTs. However, another panelist considered the recommendation misleading because, in order for propensity scoring to produce this equivalence, it would be necessary for all covariates to be measured perfectly, all relevant covariates to be included, and none of the covariates to be mediators or proxies for the independent variable of interest. In his opinion, these assumptions are impossible to meet are rarely well-approximated. A third panelist pointed out that propensity score analysis controls only for initial differences that are anticipated, measured and observed, while RCTs also control for unknown differences. In addition, the unobserved heterogeneity in error terms is not captured by propensity score analysis. In short, there was a concern that there is still a lot of art to the science of propensity scoring.

Other concerns were expressed about the preference for cross-sectional survey designs. One panelist felt that if the recommendation implied *rolling* cross-sections[8] and time-series analyses, "then you have a case" for recommending cross-sectional surveys. Another expert warned that although a repeated cross-sectional design eliminates attrition and the effects of familiarity with the questions, it sacrifices the explanatory power that panel studies provide by measuring mechanisms of change over time within individuals.

<u>Evaluating a "Moving Target"</u>

Often, after a communication campaign or intervention is launched, process data suggest the need to revise the program. Although revisions should be minimized if the program is an attempt to replicate an efficacious model, some program corrections will probably

[6] Evidence of widespread exposure is one possible addition to this list (Hornik, 2001).
[7] In this context, the propensity score is the conditional probability of being exposed to the communication program given a series of observable variables; the score has been used in both cross-sectional and panel studies that did not employ experimental designs.
[8] Re-interviews with individuals who responded to cross-sectional surveys.

be necessary and must be reflected by changes in related exposure measures on evaluation instruments. In addition, an evaluator should document each change and explain its impetus.

The typical logic model can accommodate substantial change in intervention materials and procedures. In fact, an evaluation design can probably weather the addition or elimination of entire intervention components if (1) good process data are collected and (2) intermediate or short-term outcomes are tracked through regularly fielded surveys. Ideally, repeated cross-sectional surveys should be used in tandem with pre-post surveys of a longitudinal panel because the cross-sectional surveys offer flexibility and the panel can show that exposure preceded behavior change. A rolling cross-sectional design can serve a similar purpose. Multi-level growth modeling with four or five waves of data fosters an understanding of the relationship between individual-level and community-level factors in behavior change.

Regardless of the design employed, making post-launch changes in the outcome measures employed at baseline can hamper the detection and explanation of program effects. If desired program outcomes cannot be specified at the earliest stage (e.g., if they depend on political processes that have not yet played out), one course of action is to include all likely outcomes in the logic model and baseline measures. As the program takes shape, outcome measures that prove to be tangential can be dropped, making space to strengthen core measures. Having some baseline anchor for the outcome of eventual interest may be preferable to the alternative: using a stripped down survey at baseline and adding measures to subsequent survey rounds as program objectives are clarified.

From a lifespan perspective, the relationships between many behaviors and their antecedents change over time. Adopting such a perspective may enable an evaluator to anticipate such changes and build them into logic models and measures.

Investigating Secular Trends

Secular trends -- changes over time that are not due to the program -- can be observed not only in health-relevant outcomes (e.g., the formation of positive habits) but also in the contexts of these outcomes (e.g., access to digital media). An evaluator making the claim that change in a health indicator was brought about by a community-wide or national intervention must be able to rule out secular trends as an alternate hypothesis.[9]

The direction and magnitude of ongoing trends should be taken into account; a program-induced change is marked by a significant change in slope in a trend. A good way to refute a charge that a secular trend was wholly responsible for an observed behavior change may be to employ a "switch-back" design. Experimental and control conditions are switched after more change is observed in the experimental community than in the

[9] Even randomized studies of only two communities are subject to this sort of bias. They rest on the assumption that the communities were equivalent at the outset and would have changed at the same rate, an untestable assumption (Hornick, 2001). Drawing a larger sample of communities enhances generalizability.

control community. If change levels off in the new control site and accelerates in the new experimental site, then attributing cause solely to a secular trend can be ruled out.

Studying the factors that drive secular trends can help an evaluator explain how campaign components function. For example, public policy can both affect secular trends and be affected by them, a dynamic that should be considered in evaluating an agenda-setting intervention.

Measuring Campaign Effects at the Supra-Individual Level

Many mass communication efforts not only affect individuals but also reverberate at multiple levels of the socio-political environment. Logic models should specify the pathways through which changes at higher levels of the system are expected to be tied to individual-level changes (e.g., school adoption of a drug prevention program might be expected to lower student drug trial). Caution should be exercised in direct attribution of higher-order effects to communication campaigns.

Multi-level models (e.g., HLM and random coefficient models) can combine data about individuals and their environments in the same multi-level analysis, but the utility of this analytic method is limited by the adequacy of the measures of the higher-order constructs. Higher-order constructs have been measured in a variety of ways, but each has its drawbacks (see table 2).

Table 2. Strategies and associated drawbacks for measuring higher-level constructs

STRATEGY	DRAWBACK
Aggregate individual measures into a higher level construct. For example, measure the satisfaction with the availability of health information from an organization among a sample of individuals and use the average score as a measure of the accessibility of health information at that organization.	Findings from individuals can be generalized statistically only to other individuals in the population from which the study participants were drawn. Here, the unit of analysis is the individual. Multiple data points with the organization as unit of analysis would be needed in order to generalize to other organizations.
Aggregate scores from key community informants into measures of a group attribute (e.g., community readiness to change [Thurman, Plested, Edwards, Foley & Burnside, 2003])[10]	There has been criticism of using individual behavioral constructs to explain supra-individual phenomena instead of drawing constructs from sociology, economics, and other fields with a supra-individual focus. Also see box above.
Develop scales (e.g., the Level of Institutionalization Scale [Goodman, McLeroy, Steckler & Hoyle, 1993]) and use analytic measures (e.g., social network analysis [Boulay, Storey & Sood, 2002; Rogers & Kincaid, 1981; Valente, 1995]) specific to supra-individual phenomena	It has been a challenge to find meaningful, agreed upon, easy-to-use measures for constructs such as social capital that go beyond the individual level of measurement.
Equate media content with institutional agenda-setting (e.g., Yanovitzky, 2002)	Counts of the appearance of a topic in the media have become harder to interpret now that 24-hour news channels present stories repeatedly.
Count changes in organizations (e.g., meeting frequency) or governmental policy	There is no central repository for information about local policies and no easy way to obtain data about health insurance company policies.

Insightful qualitative work has assessed higher-order constructs such as women's empowerment, but more methodological work is needed to develop valid, reliable, user-friendly quantitative measures of key supra-individual constructs. Some of the constructs that merit attention are facility quality, collaboration among institutional networks, and the community-level outcomes in the Rockefeller Foundation's Communication for Social Change Model.[11]

[10] In a study funded by ONDCP, Westat documented "buy-in" by state and organizational gatekeepers to the anti-drug campaign. Influentials were interviewed to assess institutional-level adoption or adaptation.

[11] The community-level outcomes include leadership, ownership, equity of participation, social cohesion, social norms and knowledge equity. The Social Change Model is accompanied by suggested indicators for

While this scale construction proceeds, evaluators can employ other methods to capture supra-individual variables. Methods such as: (1) mining archival records for advocacy, policy agenda and funding patterns, (2) qualitative analysis, and (3) tracking commodity availability (e.g., the number of safe venues for exercise) have much to offer.

Teasing Apart Effects of Exposure Channels

Large-scale communication programs often disseminate messages through a variety of channels. An understanding of the primary function of each program element would feed into logic model development for future campaigns. Similarly, if evaluators were able to attribute campaign effects to individual channels, future decisions about the allocation of campaign resources would be on much firmer ground.

However, assessing the contribution of each channel to a campaign outcome means measuring exposure to each channel accurately and that is difficult. Extensive interviews (with both open-ended and time-bound, scenario-specific questions) are necessary to capture aspects of exposure such as message recognition, recall and comprehension. Dose-response information is also needed, but channel-specific measures of exposure dosage are often unreliable. To make matters more complex, interpersonal interaction diffuses messages through a population and reinforces them indirectly. As one panelist put it, "How do you capture buzz? How do Sony, NIKE and Pepsi do it?"

If a program works by means of synergistic interaction among channels, analyzing separate channel effects could be misleading. This would be especially true if different channels target separate levels of a system and the levels are expected to come together to create an overall program effect. When synergy is predicted by the program logic model, the question becomes how channels complement and reinforce each other.

To estimate differential effectiveness, one could conduct experiments or quasi-experiments on various combinations of campaign components in different cities. Rolling cross-sectional studies might suffice if channels were deployed sequentially and exposure could be measured accurately. Meta-analyses of previous campaigns that used various combinations of channels could also be conducted. Finally, in a lagged approach, apparently effective elements of an initial campaign could be retained and tested in a later, more streamlined campaign. Ultimately, if cost data had been collected during the intervention, establishing channel effectiveness would permit estimation of relative cost-effectiveness by channel. The potential of these methods should be explored.

However, such topics may be better suited to basic research than to program evaluation because of their scope, the large sample sizes that would be required, and a thicket of methodological complications and potential confounds:

---

assessing these outcomes, and a recommendation to use community dialogue and collective action to impact the outcomes and indicators.

- Response often depends on dosage, dosage depends on investment in a channel, and investment is usually a function of expected channel reach.
- Some campaign elements are easier to remember than others, biasing effect sizes.
- Channel effectiveness can depend on audience readiness.
- Certain media can be more appropriate than others for particular messages. For example, newer media such as instant messaging that are accessed at an individual's convenience may be the best way to prompt behavioral choices normally made at certain times of day.
- Within a medium such as television, genre (e.g., drama vs. sitcom) may interact with message content.
- Some media may be more effective than others among certain audiences. They may be more credible or, as Uses and Gratifications theory would predict, more intrinsically gratifying (see Slater, 2003; Thapa, Graefe & Absher, 2002; Eveland, 1997; Blumer, 1979). For example, newer media may have more appeal and be more effective for youth than for older people.
- Small-scale experimental tests could provide some useful information, but would lack external validity and probably be insensitive to unintended consequences. For example, small, controlled studies could not readily answer questions about whether there is a saturation point with a medium like TV and whether there is a loss of credibility or a boomerang effect when this point is reached.
- A real world study of effects of various campaign components must take into account effects of similar and/or competing messages coming from other sources.

Given these challenges, the potential of qualitative methods should be mined. For example, ethnographic and diary methods can be employed with a sub-sample of the population to gain insight into channel effectiveness. Quantitative process data (e.g., records of media buys) can help to validate such qualitative findings.

Some panelists questioned the focus on channel effectiveness, arguing that channel reach is and should be the "trump" consideration, especially when reach is combined with low per-contact cost. They felt that television maximizes both reach and cost criteria in most cases in the United States. One expert added that although some topics do not lend themselves to a 30-second treatment on TV, a short TV ad has the power to set the agenda and drive people to other channels that can provide more intensive exposure. Other panelists argued that effectiveness should be primary, but conceded that "reach is often the only game in town." As a bottom line, most campaigns should not be expected to affect individual health-relevant behavior when the message is sent through a single, unreinforced channel, whatever its reach.

Panelists made nearly identical comments when asked about the advisability of trying to tease apart the contribution of communication from the impact of other kinds of program components in a multi-component intervention. They added that it can be difficult to distinguish communication aspects from the rest of a program, especially in social marketing where *product, price, place*, and *promotion* considerations overlap to some degree. Policy change can have an overriding impact that masks other effects. The few available examples of differential attribution of cause to separate components within a

single program are based on a fair amount of guesswork. In short, expectations for success in components analyses should be tempered.

Timing Exposure, Awareness, and Effects

Capturing the effects of a communication campaign can be largely a question of timing. To constitute scientifically credible evidence that a campaign was effective, findings such as positive trends in exposure to the campaign and positive associations between exposure and desired outcomes must be statistically significant. Statistical significance can hinge on timing survey waves to capture the maximum level of campaign exposure awareness and the strongest campaign effects.

There are no general rules for timing data collection, but an understanding of the factors that converge to affect the shape of exposure and outcome curves can inform an evaluator's decisions about when to begin to collect data, how many survey waves to conduct, and when to terminate data collection. The following sections of this report identify and discuss a number of the factors that should influence timing decisions.

Exposure curves. Unless a campaign is dull, or fails to address barriers to exposure properly, awareness of a campaign should build and continue as long as its components are still in place. Assuming continuous message dissemination, cumulative self-reported audience exposure usually takes an s-shaped diffusion curve. Both the slope of this campaign exposure curve and its eventual peak are affected by:

- the initial level of audience familiarity with the message,
- audience characteristics such as age and ethnicity,
- the reach and popularity of channels employed,
- message characteristics such as format and complexity,
- message repetition,
- the seasonal relevance of the message,
- audience involvement with or "shock value" of the message, and
- clutter in the communication environment.

The volume of campaign inputs (e.g., media buys) can be the key determinant of the speed of exposure and of the ultimate degree of penetration a campaign achieves. For large-scale, broadcast media efforts, noteworthy levels of campaign awareness can be almost immediate. To date, such campaigns have averaged a 40% post-campaign exposure level according to one meta-analysis (Snyder, Hamilton, Mitchell, Kiwanuka-Tondo, Fleming-Milici, & Procter, 2004).

In practice, messages may not be disseminated continuously, and recall of campaign exposure will be associated with the peaks and valleys in campaign outputs such as television ads. Decisions about scheduling outputs should be based on campaign objectives and the resources available. For example, program planners may decide to pulse outputs to keep exposure levels relatively high while conserving limited resources. If the campaign objective is to promote a behavior that must be performed repeatedly,

intense "flights" of campaign outputs that prompt behavioral trial can be followed by smaller reinforcement waves to encourage long-term maintenance of the behavior.

After the campaign ends, reported exposure often has a smoothly decelerating, negative quadratic curve of decay. The point beyond which campaign exposure is unlikely to be recalled normally ranges from one to six months post-campaign, but the use of narrative message formats such as soap operas may extend this period.

Attitudinal and behavioral curves. It is easier and quicker to create awareness of a campaign than to impact psychosocial determinants of behavior change such as attitudes. Attitude change may lag exposure by several months, and behavior change is likely to take even longer. The cumulative progression of these campaign effects often follows the same S-shaped dissemination curve that campaign awareness does, but with lower peaks.

In general, more attitude and behavior change will occur more quickly if:

- the dosage of exposure to campaign materials is high,
- messages are disseminated through multiple channels,
- the campaign conveys new information,
- campaign content is based on a valid attitude change theory that is applicable in that context,
- the behavior is easy to perform (e.g., calling a hotline),
- audience members have opportunities to perform the behavior,
- services are available to support the new behavior, and
- family and friends support it.

Some behaviors are likely to deviate from the S-shaped acquisition curve. For example, addictive behavior change often entails relapse (Prochaska et al., 1992). Other behaviors may go through a latent stage and then be primed by an event.

If heavy doses of broadcast media are used to address a new issue that is salient to a target audience, behavior change can occur quite quickly. Compared to community-based programs, national mass media campaigns may generate faster short-term effects. When a behavior is easy to perform, enjoys high perceived utility in an audience, and has no significant costs, effects may occur after just one exposure.

Even when the desired behavior is more demanding and the benefits to the audience more temporally remote, short-term results can be apparent by the end of a well-designed campaign. However, when the issue is mature and a health-compromising behavior is consistent with social norms, behavior change may require multiple campaigns over many years.

Effects tend to wane after campaigns end,[12] displaying a smoothly decelerating, negative quadratic decay pattern. If incentives are offered for behavior change during the campaign, decay may be more rapid. Also, decay can be accelerated if competitors actively endorse unhealthy choices in what one participant called "the drumbeat of commercial product promotion." Conversely, periodic reinforcement of key campaign messages and supportive institutional and policy changes can add longevity to a campaign effect.

Assuming adequate program funding and implementation, theory is the best guide to forecasting the lifespan of a campaign's effects. For example, both the Elaboration Likelihood Model (Petty & Cacioppo, 1979) and attitude accessibility models (e.g., Fazio, 1989) predict shorter-lived campaign effects when there is less audience involvement with a message.

Spacing survey waves. A program logic model should specify how long it will take to achieve program effects and survey waves should be spaced accordingly. The spacing decision should be leavened by the evaluator's experience with similar sorts of campaigns. Finally, reporting requirements (e.g., from fixed funding cycles) must be taken into account.

As previously noted, it may be advantageous to collect data in relatively small, closely spaced waves or continuously fielded surveys, rather than in two large pre/post waves. Frequent cross-sectional waves allow the evaluator to estimate population-level change in causal mechanisms, to track exposure and responses to materials so that mid-term corrections in the media mix can be made, and to capture effects that have short half-lives or unpredictable patterns (e.g., effects that depend on press coverage).

Multiple waves in a time series that begins well before the campaign can reveal the slope of secular trends in behavior, and information about prior slope makes pre-post evidence of behavior change much stronger. For the best information, the experts advised combining data from a rolling cross-sectional survey with data from a small longitudinal panel so the extent of change and the mechanisms of change can be investigated. Panel waves should be far enough apart so that responses are not biased by previous waves.

If the goal is to tease apart the separate effects of two related campaigns running in tandem,[13] survey waves should follow the launch of specific campaign elements. Similarly, bursts of broadcast media exposure may have widespread but short-lived effects, so waves should follow the media bursts closely.

[12]Some target behaviors do not return to baseline levels after a campaign. A sustained change may indicate: (a) that audience members were ready to change or were redirected towards a different developmental course by the initial change, (b) that the behavior was easy, non-seasonal or low-cost, (c) that the behavior influenced social norms, or (d) that the campaign echoed an existing secular trend.

[13] Several experts discouraged the attempt to separate the effects of simultaneous campaigns with similar messages. Instead, they recommended summing exposure to either campaign into an overall exposure index.

Of course, continual or very frequent survey data collection is expensive and may not be necessary if there are alternative sources of mid-point data (e.g., records of calls to a help line, in-market intercept tracking studies, or pertinent omnibus survey data). If a campaign is multi-year or a behavioral objective has limited seasonality (e.g., sunscreen use in summer), annual surveys combined with supplementary data should provide sufficient information.

Separating the Effects of Program Exposure and Background Noise

Campaign messages compete with "background noise" from other messages for the attention of target audiences. Background noise may undercut, confuse, distract or reinforce the campaign messages. In some cases, the noise can be influenced to amplify program messages.

The evaluator should consult multiple theories and experienced key informants to identify potential sources of noise because statistical analyses cannot control for the influence of noise factors that were not measured. Relevant background factors should be reflected in the program logic model and major ones should be measured sensitively.

With early identification of noise factors, it may be feasible to design the potentially confounding effects of noise out of a study. For example, an intervention-plus-noise condition could be compared to a noise-alone condition in an experimental or quasi-experimental design.

Even the most diligent evaluator can overlook an environmental factor that threatens to be a confounder. However, evaluation strategies can be structured to redress such an oversight early in the data collection process. Early warning of an important environmental noise factor can come from:

- repeated surveys that ask open-ended questions about exposures and can detect unexpected or non-incremental changes in outcomes
- commercial databases
- environmental scanning through key informant interviews and monitoring news media coverage, especially of stories about program-relevant policy changes

After data are collected, the effects of noise can be reduced by:

- adjusting results statistically for measured extraneous variance
- analyzing residuals and error terms to determine if unmeasured factors are undermining attribution
- using propensity scores to handle several covariates simultaneously, being sensitive to overinterpretation and the problem of reciprocal causation of exposure and outcome

Despite the most valiant efforts, an important background noise factor can be unanticipated and unmeasured. Confounded results must be acknowledged and noise factors must be taken into account when making claims on the basis of outcome data.

Reliability of Self-Reported Media Exposure

Very little formal research has been conducted to gage the reliability of self-report measures of media exposure. This area merits further research attention. A number of studies have shown that false exposure reports and other sources of bias account for only about 2-5% of self-reported message recall, but others have found "definite" false recognition rates of 10% or more, and rates of 25-30% when the response was "maybe saw." These estimates are produced by studies that ask about exposure to "ringer" or "foil" messages that were not actually sent (e.g., Southwell, Barmada, Hornik & Maklan, 2002; Slater & Kelly, 2002). Slater & Kelly (2002) suggested controlling for false recognition bias by using ringer recognition as a covariate in statistical analyses.

Even with such controls, some experts continue to question the reliability of self-reported exposure. Measures of recall are subject to systematic bias from selective attention -- the predisposition to be exposed based on an individual characteristic such as race. Moreover, many respondents report exposure to a real message through the wrong medium (e.g., that they saw on television a message sent via internet).

The reliability of self-reported exposure recall increases if:

- campaign messages are clear and easily differentiated from others
- multiple measures funnel to specific ones about particular executions
- exposure context and time frame are provided
- exposure is recent
- exposure frequency and channel details are not sought
- the respondent is motivated to comply, values the outcome variable, and finds the topic being addressed personally salient

It is also very important to pre-test survey questions. Some studies have found that reports of behavior vary depending on the wording of survey questions or response options. Survey descriptions of ads should be tested to confirm that people who have seen the ads being described recognize the descriptions easily. Pre-testing the ringer or dummy measures is also useful; ringers should be selected to be plausible but not easily mistaken for the real ads.

Asking questions about knowledge of the campaign along with multiple measures of recall can help in triangulating exposure data (see the description of the Mass Media and Health Practices project in Rice & Paisley, 1981). Interviewers and data coders should be very familiar with all campaign outputs so they will recognize respondent references to the campaign.

Researchers interested in this area should note several measurement difficulties. First, although recognition tends to be more reliable than recall, the domains are not mutually exclusive (Southwell, et al., 2002). In addition, secondary exposure (hearing about a mediated message instead of hearing or seeing it directly) maps onto and inflates

recognition. Finally, it is difficult to measure depth of exposure (e.g., the difference between mentioning a PSA to others and discussing the content of that PSA).

Where possible, exposure should be manipulated to create an objective measure with which to confirm self-reports. New web evaluation techniques should be a boon in this regard. These techniques allow the researcher to track the specific websites visited by a user, how often he or she returns, the time elapsed before clicking ahead, and other exposure dimensions.

How Important is Measuring Exposure Precisely?

Communication program evaluators debate the importance of a high degree of precision in measuring exposure. Those who feel that it is critical that exposure measures be highly reliable, sensitive and specific make the following points:

- Program effects may be undetectable without precise measures of exposure dosage.
- Exposure data are essential in modeling the change process. By themselves, outcome data from intervention and comparison communities cannot justify causal claims (e.g., because of discrepancies between planned and actual levels of message delivery in the intervention community).
- Precise exposure measures make it possible to conduct assessments of the differential effects of various media and cost-effectiveness analyses to inform plans for future media use.
- Precise, valid, continuous measures of exposure allow dose-response analyses to be conducted, contributing to the justification for causal inferences (see the Belmont recommendations, Appendix C).

Evaluators on the other side of the debate argue that measuring self-reported exposure with a high degree of precision may not be possible, and that some error in exposure estimates is acceptable. Although such error lowers effect sizes (which are already modest for campaigns), these evaluators maintain that:

- validity of exposure measures is more critical, especially in non-experimental designs with their great potential for self-selected or endogenous exposure
- other exposure data (e.g., information about geographic variations in a media buy) can compensate for the unreliability of self-reported exposure data.
- all methods are subject to bias. If a consistent method of measuring exposure is used over time, the consistent bias will not disguise significant trends in exposure.

Conducting basic research to test the reliability and validity of exposure measures prior to their use in program evaluation makes evaluation research more credible. However, measures that have not undergone formal psychometric testing can be used in field evaluations if there is no good reason to question the credibility of responses (i.e., exposure measures seem robust and there are no competing programs of note). Some analyses of reliability (e.g., Cronbach's alpha) and validity (e.g., construct validity) can be conducted post hoc with data collected during the evaluation. Reliability analyses of

attitudes and exposure are especially important because low reliability in either measure can mean underestimating the correlation between them and hence the program's effects.

Using Indicator Data to Evaluate Programs in the United States

In the international evaluation arena, the term *indicator data* often refers to information collected by governments and non-governmental organizations to inform health service management and policy making. Indicator variables have stable definitions over time and are collected or summarized at regular intervals in surveys, clinic reports, etc. Indicators can measure health behaviors and their antecedents, the incidence and prevalence of diseases and conditions, and service access and utilization. They are designed not to evaluate the outcomes of individual programs, but to reflect the combined impact of all the activities being conducted to reach a health goal (UNAIDS, 2000).

In the U.S., evaluations of campaigns and programs at all levels of reach may stand to benefit from the use of indicator data. Indicators can chart potentially confounding secular trends and provide baseline information. Indicators can also serve as proxy measures of program outcomes if (a) they *should* be sensitive to the effects of that intervention according to a logic model, and (b) there are no other influences that could explain changes in indicator trends. To explore linkages between program exposure and longer-term outcomes of campaigns, indicator data can be juxtaposed with program monitoring data, media buy records, media tracking reports, and meter or diary-based ratings of viewership, readership or listenership (e.g., from Neilsen or Arbitron). Thus, although imperfectly suited to the evaluation needs of most programs, indicator data can complement other evaluation procedures and stand in for custom data collection when evaluation resources are scarce.

The challenge is finding good time-series data that are sufficiently relevant to specific campaign objectives. Even when behavioral indicators are relevant, standard indicators rarely include measures of psychosocial indicators of behavior change or campaign exposure. However, there are many instances in which campaign-specific exposure questions have been added to standard indicator lists for a specified period of time. Similarly, for a fee, a few indicator collection systems (e.g., Porter-Novelli's "styles" surveys) will administer a wide range of item types along with the usual questions asked.

One evaluation expert argued that indicator data are not worthwhile as a campaign evaluation tool, but are useful only as a "news hook" to put an issue on the public agenda and win the investment of opinion leaders. Another panelist agreed that indicator data have an agenda-setting function, but considered them potentially valuable in multi-level models of program effects. Here a large-scale program would be hypothesized to spark media coverage and public debate which, in turn, would be expected to affect outcomes of interest (see Stryker, 2003). This kind of application might combine state-by-state indicators such as the Behavioral Risk Factor Surveillance System (BRFSS) with data from news content analyses, program monitoring, or media buy records.

Other panelists pointed to the value of data like the BRFSS in uncovering health disparities across states, detecting synergies in related behaviors (e.g., diet gains from exercise messages), triangulating non-experimental evaluation approaches, and confirming an evaluator's own estimates from customized surveys. To make indicator even more useful in program evaluation in the U.S., panel members suggested:

(1) adding more theory-based indicators (e.g., attitudes) to illuminate causal pathways, an expansion the international Demographic Health Survey is planning,
(2) supplementing clinic data with exit interviews (conducted before, during and after a campaign) that ask clinic clients how they knew about the clinic service,
(3) using communication-specific data sets such as NCI's Health Information National Trends Survey,
(4) piggy-backing on well-conducted data collection activities such as the monthly Current Population Survey conducted by the Bureau of Labor Statistics,
(5) obtaining detailed information about indicator sampling frames and collection cycles so that single-program studies can use complementary procedures, and
(6) using the same interviewing method in customized surveys that the indicator uses because mail, web, telephone, and face-to-face data methods can yield very different results on sensitive items.

When to Lay Evaluation Efforts to Rest

If the full range of diversity in a target audience is represented in focus groups, the current standard for terminating focus group data collection is to stop holding discussions after the saturation point (i.e., when most of the ideas brought up have been mentioned in previous groups). For quantitative outcome evaluation, however, there is no clear guideline for determining that data collection has gone on long enough.

In general, the duration of an evaluation should depend on its goals, but the goals of various evaluation partners can differ. A program provider or stakeholder may be satisfied with evidence that the program reached its behavioral objective, while evaluators want to understand the programmatic and other factors that brought about the behavior change. To understand the mechanisms involved in behavior change, a campaign's logic model should dictate when to terminate data collection efforts; it may be necessary to extend data collection until program effects trail off or "boomerang."

A longer timeframe can also permit examination of "macro" outcomes. For example, if initial results show that a communication strategy is working as intended, subsequent phases of the investigation can follow related institutional changes such as the adoption of the strategy by other providers. Also, because multiyear communication campaigns evolve almost by definition, the list of unanswered evaluation questions keeps growing.

Long-term follow-up data should be collected for at least six months in most cases, but if a campaign continues to introduce new content that interested parties are keen to assess, longer-term evaluation efforts can be justified. Conversely, follow-up data collection should be discontinued and resources placed elsewhere when (1) an evaluator can extrapolate the behavior curve to extinction or maintenance, (2) an environmental change

occurs that could account for continued effects, (3) attrition rates climb too high, (4) donors lose interest and/or the money runs out ("no one is listening"), or (5) a program can gain sufficient information from routine monitoring.

Although relatively long data collection timeframes may be "best practice" from an evaluator's point of view, decisions about terminating data collection are often made on the basis of budgetary factors. If there were formal guidelines for the length of data collection periods in evaluations of mass communication, such guidelines might help persuade funders to invest in longer data collection periods.[14]

Final Comments on Design

One panelist commented that "…there is no single set of standards for evaluating a category as diverse as full coverage campaigns." Another argued that the bottom line of this report should be an admonition to avoid tailoring interventions to meet research design criteria, and to pursue the development of research designs that are appropriate for the phenomena of interest – full coverage programs. Of course, the most rigorous designs and/or statistical procedures may not be merited by the budget, timeframe or nature of a given intervention. Moreover, in some situations, the preferred design may be the one that produces the cleanest, most intuitively plausible evidence for funders and other stakeholders. For example, the power of the telling anecdote in testimony to policy makers has been demonstrated repeatedly.

It should be noted, however, that when programs lack political support, qualitative data tend to be disparaged and more stringent levels of proof demanded. Experience has shown that if researchers reach controversial conclusions, they can expect fierce attacks and must display technical wizardry to defend their findings. In general, designs that triangulate quantitative and qualitative data are best accepted; in most circumstances, they are worth the additional expense. Large, multi-year campaigns may merit the integrated use of multiple quantitative evaluation designs, each of which addresses different inference issues.

Methods: Statistical Issues

Statistical issues were emphasized in the earlier Belmont Report but were largely beyond the scope of this discussion. However, a few general recommendations about statistical analysis were made by the CDC expert panelists and they dealt at some length with the use of imputation to cope with missing data.

General comments

The stronger an evaluation's design and measures, the less need there is for elaborate, cutting-edge statistical analysis. However, as the previous section of this report detailed,

[14] In developing such guidelines, the criteria for deciding that an appropriate level of evidence has been achieved offered in *The Community Guide for Preventive Services* (www.thecommunityguide.org) should be considered.

the designs typically used to evaluate large-scale communication campaigns and programs (even some of those that employ randomization) have numerous threats to internal validity. Although they are no substitute for theoretical coherence and good measurement, several statistical techniques can help to counter threats to internal validity that were not dealt with at the design or data collection stages.

The panelists addressed the need to account for (a) design effects (e.g., the effects of using convenience samples) in survey data and (b) for loss to follow-up in panel studies. In addition, having a large enough sample size to provide adequate statistical power to detect campaign effects was described as critical in health communication research. As elsewhere in the discussion, the experts called for attention to the level of analysis of each variable in a model to ensure that HLM methods were utilized when necessary. One expert recommended Discrete Factor Methods (see a paper by Tom Mroz at http://www.unc.edu/~mroz/papers/dfmnew/dfmnew5.html) for dealing with joint endogeneity of outcomes in complex models. Another suggested that cross-lagged Structural Equation Models can detect program effects in longitudinal data while controlling for reciprocal/reverse causality. Bayesian statistics were suggested as a way to deal with various degrees of credibility of responses and sensitivity of measures.

There was a longer discussion of statistical modeling of competing theories (see the related section on design issues on page 17). In most program evaluation, statistical tests are conducted to assess program outcomes, not theoretical integrity. However, a program's logic model should be theory-based at least in part, and outcome measures should cover the main theory or theories that informed the design of the intervention.

Some experts felt that, unless the original intent of a program is to test a theory, statistical modeling of competing theories is an academic exercise -- a nice bonus, but not the main point. Seldom does an evaluation instrument measure enough of the variables from multiple contending theoretical frameworks to permit the theories to be modeled separately. A limited number of questions can be asked, and they are usually based on either a single theory or on portions of various theories. In the rare instances in which theoretical integrity is preserved for more than one theory, path analysis or Structural Equation Models can trace the theoretical pathways with the greatest explanatory power.

Finally, full reporting of statistics (e.g., means, variances, effect sizes), sample sizes, and sample descriptions was described as critical to transparency and scientific progress. In practice, this important information is often omitted from evaluation reports and published articles. The length restrictions imposed by public health journals are a barrier to reporting complex analyses fully.

Imputation

In most data sets, some variables have missing values. Data gaps occur when respondents drop out of a study voluntarily, are unable to continue participation, refuse to answer some questions, or provide answers that cannot be interpreted or scored. When

data are missing, there is a loss of statistical power and a risk of bias in a sample designed to have certain characteristics.

The problem is particularly acute in cohort studies. When there is attrition after the initial round(s) of a cohort study, power cannot be maintained by replacing respondents. In addition, dropouts may differ in some systematic way from individuals who remain in the panel, so their loss introduces selection bias.

There are several ways to deal with missing data. They can be ignored, or cases with missing data can be deleted from the sample (so-called listwise deletion). Alternatively, subsamples can be weighted[15] to correct for attrition. Finally, the problem can be addressed by imputation, the estimation of missing data within individual cases on the basis of other information.

Imputation adds noise to a dataset and can flatten campaign effects artificially. This cost should be weighed against the selection bias that can be introduced by listwise deletion (i.e., dropping cases with missing data). Roth (1994), Little & Rubin (1987) and Wothke (1998) concluded that imputation is superior to listwise deletion, but a simulation study comparing two statistical procedures for employing a state-of-the-art imputation method found that one procedure introduced more bias than listwise deletion (Allison, 2000).

The decision to impute must be made on a case-by-case basis. One panelist suggested creating a sample composed of the cases that would remain after listwise deletion, comparing that sample to the full sample on items with no missing data and, if the results of these analyses differ, making imputational adjustments for items with missing data. In this circumstance and all others, imputation should be kept to a minimum. No more than 5% of the cases should be imputed, 2% per cent is a preferable cutoff point, and imputation of campaign exposure or outcomes should be avoided altogether.

One panelist considered imputing ideational or behavioral variables more problematic than imputing demographic variables. Another panelist suggested creating variables by combining items that do and do not have missing data; the resulting values would reflect some actual information from the respondent. Several panelists said it was standard practice to base degrees of freedom on the number of complete cases so as not to inflate power, and then to impute data to limit selection bias.

The first step in imputing data is to test the *missing at random* (MAR)[16] assumption (Little & Rubin, 1987). If the pattern of "missingness" is random, then one of several

[15] Multiplying each case in a group (e.g., low income individuals) by a number that makes the group's proportion in the sample equal to some desired proportion.

[16] For data that are *missing completely at random* (MCAR), the probability that an observation is missing is unrelated to the value of that variable or other variables. For example, missing data on family income would be MCAR if people with low incomes were no less likely to report income than people with high incomes. The more typically valid MAR assumption requires only that there be no association between the probability that a data point is missing and the value of that data point *after* controlling for another variable. For example, reported income may be related to depression and to the probability that income is reported.

imputation approaches can be used (Roth, 1994). If missingness is not random, the missing cases are termed *nonignorable* and the evaluator must rely on emerging procedures that do not make the MAR assumption, such as Heckman models and pattern mixture models (Hedeker & Gibbons, 1997).

Evaluators should specify the method of imputation for MAR data because some methods are better regarded than others.[17] For example, substituting the mean response for a missing value is considered inferior to methods such as multiple imputation (MI); MI decreases variability in the data set and increases the risk of type 1 error. MI generates multiple values for a missing data point by Monte Carlo simulation, and then integrates them into an overall confidence interval through a maximum likelihood procedure. Another algorithm for imputation, the expectation maximization (EM) method, uses values from the dataset as the basis for estimates. EM has been criticized for introducing bias because it does not use all the available information. "Raw" or "full information" maximum likelihood methods are being developed to address this shortcoming.

If the results of key analyses differ with and without imputed data, both sets of findings should be reported and the 2% rule for imputation applied. Outcomes detected only when data are imputed lack credibility. However, if bottom-line findings are not sensitive to the presence or absence of missing data, and the imputed information adds confidence that the population is fairly represented, then one has a strong argument for relying most heavily on the findings that include imputation. The argument can be further strengthened by citing literature in support of the practice (Allison, 2000; Roth, 1994; Little & Rubin, 1987; Wothke, 1998).

The best advice is that imputation be kept to a minimum by making a heavy initial investment in the data collection process. For example, careful questionnaire development can minimize the occurrence of missing data by detecting and eliminating item order effects, reducing response burden, and identifying and neutralizing sensitive questions. It may also be prudent to select a type of data analysis that permits available information from a case to be utilized even if the respondent drops out before a study ends (e.g., event history analysis; survival analysis, or hierarchical linear modeling).

III. RESOURCE-SCARCE CIRCUMSTANCES

Scarcity of evaluation resources can restrict an evaluator's choice of study design and method and force the neglect of interesting questions about the execution and outcomes of a communication program. When evaluation funds are scarce, evaluation should be limited in focus to the highest priority questions; rank-ordering program aims and related

---

If the association between income level and income reporting is no longer significant when depression is held constant statistically, then the data would be considered MAR.

[17]General FAQ #25: Handling missing or incomplete data. This clear summary from the University of Texas reviews various methods of imputation and the statistical software for employing them. It is available on the web at www.utexas.edu/its/rc/answers/general/gen25.html.

research questions on the basis of a theory-derived logic model is considered best practice. However, if major aspects of the intervention are already understood and program implementation is tracked routinely, the evaluator may be well-advised to illuminate the weakest link, or to focus on the concerns of the most powerful and contentious critics.

If the program itself is severely underfunded, it may be best to redirect outcome evaluation funds to program activities. Programs should not be expected to achieve objectives specified in a logic model that fails to take financial practicalities into account. The recommendations below assume that programs are sufficiently well-funded to merit some evaluation.

Designs for Resource-Scarce Circumstances

Although maintaining contact with respondents over time is costly, longitudinal panel studies can be less expensive than cross-sectional surveys because fewer survey rounds are needed to detect change of a given magnitude in an outcome of interest. However, the most cost-effective survey evidence is probably a cross-sectional survey with propensity score analysis and appropriate tests for endogeneity. One panelist volunteered the opinion that a post-only survey of this type is fairly credible.

Unfortunately, many programs lack the resources to construct and survey samples large enough to detect effects statistically, regardless of the survey design. In such cases, an outcome evaluation can rely on alternative methods of data collection including:

- unobtrusive proxy measures (e.g., condom sales instead of reported condom use),
- routinely collected indicators that map closely to key campaign outcomes,[18]
- campaign-specific items (e.g., exposure items) "piggy backed" on an existing surveillance system,
- surveys of members of an ongoing, commercially maintained Internet panel,[19]
- textual analysis of reports, clinical records, publications, and media coverage
- in-depth interviews with key respondents and direct observation to find out what people are actually exposed to and how they are reacting.

Choosing to Limit Outcome Evaluation

Even when these alternative methods of outcome (or summative) evaluation cost a lot, they can still fail to provide clarity about the pathways or mechanisms of change. Instead of conducting a summative exercise, it is wise to allocate a program's evaluation resources to process evaluation when:

- outcome data would be uninterpretable because there have been deviations from the intervention design and their nature and/or extent is unclear

[18] Care must be taken in interpreting archival indicators; a campaign can be associated with an increase in a problem indicator by boosting awareness, care-seeking, and subsequent diagnoses (e.g., skin cancer screening). See page 23 for additional discussion of indicators.

[19] While Internet panels may not be fully representative, their bias is likely to be consistent over time, so they may still be useful.

- the design holds little promise to begin with
- there is insufficient time to conduct a meaningful summative evaluation (e.g., in an emergency response, or when a logic model predicts a time lag between intervention and outcomes that exceeds the research funding period)
- outcome behaviors are adequately monitored through outside mechanisms
- a model with documented effectiveness is being replicated with target audiences and contexts that are reasonably similar
- stakeholders have privacy or other ethical concerns about outcome data collection
- the organizational activity is routine (e.g., ongoing media relations)

Process evaluation, on the other hand, is essential in all cases, both for fiscal accountability and for alerting program staff to the need for program adjustments.[20] In addition, process data that show that an obviously worthwhile program is reaching more and more of its intended audience over time can amount to evidence of program effectiveness.

IV. Addressing Racial/Ethnic Health Disparities

Racial and ethnic minority group members in the US suffer disproportionately negative outcomes in many health areas (Smedley, Stith & Nelson, 2003). Broad systems changes are needed to reduce these disparities, but health communication campaigns and programs can make a contribution. Assessing their differential impact on minority and general populations can inform later resource allocation.

If a communication program targets minority audiences, members of the audience should be involved in program and evaluation planning from the outset. The expert evaluation panel also urged paying special attention to theory selection, dedicating resources to formative research and cognitive testing, using special means to track message dissemination, and employing adequate subgroup sample sizes. These recommendations are developed in the following sections of this report.

Theory, Formative Research and Diverse Audiences

The theories that are most commonly used in communication research are fairly robust and have been validated in a variety of contexts. However, the standard meanings of and relationships among theoretical constructs can be a poor fit with the realities of vulnerable populations. Some would argue for studying the predictive validity of a standard theory with each minority audience, but such studies can siphon off resources from the intervention, sacrifice credibility with the target audience, and waste time.

In most cases, instead of conducting a predictive validity study, the evaluator should ask whether communication variables such as beliefs, access to services, and channel use are differentially distributed or fail to behave among diverse audiences as standard theory

[20] A key process question would be whether the model is being faithfully replicated.

would predict. Relevant evidence can come from previously conducted population-wide surveys that can be disaggregated into minority subsamples. Community-level barriers that are of negligible importance for majority audiences can become glaringly important for groups with the greatest negative health disparities. In light of such evidence, logic models and measures should be adjusted. Supplemental theories (e.g., theories of cultural assimilation) may provide guidance in amending a logic model that was originally based on a generic theory.

If there is an apparently satisfactory fit between theory and audience, the evaluated intervention itself can serve as a test of the predictive validity of the theory. In the event that the intervention proves unsuccessful, an evaluator who has included a comparison group from another population and thorough process measures may be able to distinguish between theory failure and implementation failure.

Even when a standard theory is adequate for use with a disparately affected audience, it is usually necessary to tailor the strategies, materials and measures that have been used with other audiences. The formative research necessary to this tailoring can also be used to refine the logic model. Evaluators should make every effort to recruit formative research participants who are truly representative of hard-to-reach audiences. Care should be taken to identify homogeneous subgroups within the minority population (e.g., by asking focus group participants which group they feel their answers represent).

Minority focus group participants should be asked whether they have concerns that psychological stress or discrimination in employment, health care, or social life will result from adhering to a recommended practice. Other barriers to compliance should be fully probed as well, particularly when respondents are from low-income populations or have relatively little power in society. Some populations are not forthcoming in focus groups, and cultural differences in interpreting a facilitator's questions can mask beliefs. Researchers should explore new ways to gather such information.

Finally, to avoid unintended consequences, some additional exploratory discussions should be held with audience gatekeepers and opinion leaders from the population under study. In the past, the best of intentions have aroused anger or even been counterproductive because salient differences among audiences were overlooked.

<u>Cognitive Testing and Copy Testing with Diverse Audiences</u>

The instruments used to measure the effects of an intervention can contain items that are unclear or have non-standard meanings to members of disparately affected audiences.[21] Conducting "cognitive tests" of the instruments prior to their use in evaluation can clarify what questions mean to members of various groups (Willis, 1999); this is always good practice. In research with minority audiences, omitting cognitive tests can result in

[21] For example, in one minority study that used the Theory of Reasoned Action (Ajzen & Fishbein, 1980), the *intention* variable had a connotation that was closer to "nice idea but probably won't do it" than to "plan to do it."

extremely questionable data. Cognitive testing is particularly important when measures are new; it is critical when items are complex or language fluency is an issue.

Cognitive tests conducted by skilled and carefully debriefed interviewers can pick up many issues of consequence. If one is conducting a general "cultural appropriateness" test on straightforward items, respondents may be tested in an efficient group setting. One-on-one interviews are probably needed when more in-depth probing is indicated.

After campaign messages and materials have been developed, copy testing (see NCI, 2003, page 213) [22] should probe on language appropriateness. Although most questions asked in copy testing with minorities should be comparable to those used with other populations under study, it is important to pinpoint the terms, usages, formats and spokespersons that resonate with minority group members. Some commercial marketing firms actually administer in-home copy tests of ads inserted into Spanish-language TV programs because it is not enough to be "culturally sensitive" in creating materials and messages – messages have to be "dead on" to establish credibility.

Measuring Message Dissemination among Marginalized and Minority Communities

The challenge of measuring message dissemination among minority audiences begins with audience segmentation. Evaluators should consider supplementing race and ethnicity information with segmentation-relevant variables such as SES, acculturation level, and social psychological orientation that will make it possible to hone in sharply on at-risk target groups.

There are three major strategies for tracking message exposure. The first is to survey a sample of the target audience about exposure to messages and campaign components. The second is to track mass media coverage or other intervention "noise" to which the audience might have been exposed. The third is to observe exposure directly or through electronic means such as label scanning. Each presents special challenges for the researcher interested in minority audiences.

Random sample surveys are particularly meaningful in that they can estimate recall as well as likely exposure. However, unless minority populations are geographically concentrated, it can be difficult and costly to identify adequate numbers of individuals eligible to participate in a survey. List-based samples for some audiences can make surveys more efficient; Spanish-surname lists for purposive sampling can be bought from commercial firms (e.g., Survey Research, Inc.). Response rates are less well known for minority subgroups than for general audiences, and study success can hinge on high rates of participation by identified eligibles. Partnering with a research group or other respected organization from within a local minority community may enhance credibility and participation rates, particularly in face-to-face research.

If a program is national, but a national survey is not feasible, surveys in sentinel cities with large minority populations may be adequate. An option that can be more affordable

[22] See also "Positioning Advertising Copy Testing" in the Journal of Advertising, 1982, vol. 11(4), 4-29.

than building a study-specific national sample is to draw a sample from a pre-existing, nationally representative panel that has members from the population of interest. Panel members have already agreed to respond to surveys in exchange for some incentive and although panel data typically suffer a lack of representativeness, they can be weighted to compensate for that bias. Finally, web access is growing dramatically for all populations, and more than 95 percent of public libraries now have public Internet access, so surveys of visitors to some web sites may soon be adequately generalizable to minority groups.

Indirect inference of exposure is possible by tracking specific content in the mainstream and minority news media several via internet search engines and some commercial vendors (e.g., Nexis/Lexis). A less expensive strategy is tracking the circulation of in-language or in-culture publications. One might also content-analyze local minority news outlets (e.g., local weekly newspapers in Latino farming communities) for references to campaigns and the inclusion of campaign messages. Monitoring the distribution of campaign-related materials and attendance at local events can also yield useful and affordable information, as can conducting intercept surveys in sentinel service sites or other places frequented by the audience of interest. Surveying clinic staff can offer inexpensive media tracking information if providers are willing and able to report patient references to campaign messages or campaign-relevant health behaviors. Finally, community leaders can be asked to describe local media consumption patterns and to identify social networks that can provide study participants.

If it is not possible to conduct a program-specific survey, the evaluator can turn to geographically coded, commercially available marketing databases, some of which rely on scan data. It may be possible to infer exposure from their extensive information about media habits and product consumption, and some commercial and social omnibus surveys will add questions on exposure or other topics for a fee. Some of the commercial marketing databases now cover Hispanic audiences but they are expensive and their reliability with Hispanics has been questioned.

More research should be conducted on efficient and effective ways to monitor the dissemination of health campaign messages to members of minority groups that suffer negative health disparities. This research should be informed by commercial marketing approaches when possible.

Special Analysis Considerations for Multiple Minority Sub-Samples

To obtain representative audience data, samples should be (a) stratified by audience segment, and (b) large enough to provide adequate statistical power to compare and draw conclusions about each segment of interest (via oversampling subgroups if needed). After data are collected, homogeneity tests should be run to determine whether differences between groups are significant, or whether the analysis can be done with a pooled sample. Testing for interactions (e.g., race x exposure) is also advisable; it can reveal the distinctions between audience segments that should be preserved or controlled for statistically in later analyses. Heterogeneity may support higher level analyses and

interpretations. In the absence of statistically significant heterogeneity and interactions, analyses should employ pooled data weighted by sub-population.

Qualitative analysis of answers to open-ended questions also requires sensitivity to differences in within-group perspectives. For example, there may be considerable heterogeneity in beliefs within a Native American sample if members of more than one tribe are included. For Latinos, country of origin can affect the meanings of language and the cultural contexts of behavior.

When subdividing minority audience segments further, one must be realistic about the scope of analysis that can be conducted. While whole ethnic groups are rarely homogenous, an over-attribution of differences to culture and an under-emphasis on similarities across cultural groups can lead to programming decisions that waste resources.

Involvement of Members of Diverse Communities

It is ideal for the evaluator to be a member of the target minority group; it is essential to work closely with group representatives throughout the evaluation process. Group membership means being alert to communication and contextual nuances that non-native speakers do not detect. For example, with "insider" assistance, an evaluator may be able to describe the complex interplay of assimilation and cultural adhesion, non-dichotomous states that fluctuate in a manner that is hard to capture. Community representatives can help to make survey questions culturally relevant, ensure the cultural appropriateness of project components and boost community participation in the evaluation.

Minority community gatekeepers and influentials should be consulted during the formative stages of an evaluation, and given periodic progress reports. Initiating and maintaining a true collaboration with these stakeholders will increase the likelihood that they will represent the campaign favorably to policy-makers at various levels. When preliminary findings are available, they should be presented to community stakeholders first. Their assessment of the validity of the findings should be sought, their recommendations for the program included in the final evaluation report, and their preferences about distribution of the report respected. They have unique insight into the impact of creating a target audience demand for services that may not be met in a community that is already shouldering more than its share of the burden of disease.

V. Policy Issues

In this report, "policy" refers to the direction and priorities of Federal support for basic and applied research in health communication. The panelists saw a lot of room for improvement in this area; one commented that Federal agencies often use an implicit logic model that predicts behavior change from the formation of a coalition and the broadcast of a public service announcement. The panelists: (1) articulated a need for more efficacy trials before taking campaigns nationwide, (2) recommended extending the knowledge base for the intuitively appealing practice of "message bundling" prior to

launching national or large-scale campaigns that rely on this strategy, (3) called for centralized identification of secular trends, (4) listed other specific areas in which communication research or research guidance is needed, and (4) recommended ways to obtain additional expert advise on evaluating large health communication programs and campaigns.

Conducting Efficacy Trials

A public sector decision to fund a national campaign should rest, in large part, on positive results from a smaller, carefully evaluated efficacy trial, just as the commercial sector examines the performance of a product or promotion in a test market prior to national roll-out. Whether an efficacy study uses an experimental design with control sites or a quasi-experimental design with comparison sites, it eliminates some of the causal uncertainties inherent in evaluation at the national level. Efficacy trials can double as pilot tests, revealing ways to improve a basically successful approach. They can mobilize leadership around a problem and, by identifying blind alleys, accelerate the larger campaign. An efficacy study does not eliminate the need to evaluate a national program, but it saves time and money in the long run.

If relevant empirical work has been done previously, a careful analysis of the earlier findings may be a reasonable substitute for an efficacy trial. However, when an intervention approach is new, a topic is politically sensitive, or an intervention is "high stakes," omitting an efficacy test could be viewed as irresponsible.

If conducting an efficacy trial is not possible, qualitative methods should be used to pre-test messages with the target audience, examining message clarity, acceptability, relevance, and other issues of potential consequence to the outcome of the intervention. There have been reports of discrepancies between focus group recommendations and the results of field-based experiments (e.g., Foley & Pechmann, 2004) and this issue deserves further exploration; a monitoring system should also be established to track audience reactions to a campaign when no efficacy test has been conducted. Such a system should provide an opportunity to correct or clarify messages well before post-test data are collected at the end of an evaluation cycle.

Studying Message Bundling

While one has the ear of an audience member, why not deliver several health messages that he or she may need to hear? Does "bundling" messages in this way risk confusing audience members or watering down the impact of individual messages?

Message bundling seems efficient and has intuitive appeal to policy makers, but there is little empirical support for the practice (Fridinger & Kirby, 2002). In the 1980's, several community-wide programs disseminated multiple behavioral messages (e.g., watch your diet *and* exercise *and* stop smoking) related to one health issue -- heart health (e.g., Fortmann, Taylor, Flora & Jatulis, 1993). In the main, the results of these ambitious projects were disappointing compared with those of single-message interventions

Another form of bundling is sending messages about more than one disease to a single target audience. The example of combining prevention messages about cervical and breast cancer has been studied and results were mixed. Whether presented alone or bundled, the mammogram message had a stable effect size, but Pap promotions may be more effective when presented singly. This limited information indicates that bundling interacts with message content, but a real understanding of the nature of the interaction would require much more exploration.

Health messages tend to be most effective when they are repeated frequently, easy to understand, and easy to act on. A bundle of messages is unlikely to have this level of accessibility, but could still influence an audience that was highly motivated to pay attention. If several messages were grouped under a unified lifestyle "brand," and audience members came to identify with this overall brand, they might be motivated to incorporate multiple behavioral "brand attributes" into their repertoires. For example, a woman might stop drinking, take folic acid, and get prenatal care to help achieve the overall brand identity of "A Good Mom from the Start." Before such an approach is taken on a broad scale, however, message bundling should be studied systematically because its potential to lower demands on the health communication infrastructure may be offset by losses in message effectiveness.

Formal outcome research using large sample sizes and various types of messages is needed to assess the independent and synergistic effects of bundled messages. These investigations will require long, in-depth surveys, the complexity of which will increase exponentially as messages about additional behaviors and diseases are introduced.

In the past, disease-specific research funding categories have been a barrier to conducting research on health message bundling, but some relevant data will be published soon (e.g., in the USDA 2005 dietary guidelines). Programs such as Steps to a Healthier U.S. are preparing to conduct evaluation studies that shed light on messages bundling. While we wait for these data to become available, we can begin to understand the ways in which messages interact by:

- reviewing evaluations of pediatric well-child efforts
- examining school-based prevention programs that teach a common set of life skills to prevent substance use, sexual risk behaviors and other risks
- tracking the literature on interactive applications (i.e., e-health tools) that combine multiple messages
- pre-testing all messages slated for bundling to learn at least something about how they interact.

<u>Centralizing Research on Secular Trends</u>

Studying secular trends through secondary analyses of survey or surveillance data is an efficient use of communication evaluation funds. Some annual surveys (e.g., Healthstyles conducted by Porter/Novelli and those conducted by the Food Marketing Institute) were designed with such public health uses in mind. Because access to these

databases can be costly, information about secular trends should be disseminated through publicly accessible channels.

There are significant gaps in the secular trend information that is tracked at present. With additional funding, ongoing behavioral tracking studies could be extended to include more communication-relevant variables. Surveys such as the BRFS, Monitoring the Future, and the Youth Risk Behavior Survey that generate state-specific data could be particularly useful in evaluating localized communication efforts.

In addition, evaluations of long-running communication programs should be funded and required to generate trend data as part of their longitudinal designs. In both intervention and control/comparison communities, they should track not only the outcomes targeted by their campaigns, but also national and local events that pertain to campaign objectives.

Final Suggestions for Advancing the Field

The Institute of Medicine report on Health Communication (IOM, 2002) did not focus on evaluation; a follow-up study that addressed evaluation would be worth supporting. The Legacy Foundation has funded an IOM report on tobacco programs that will include a discussion of media programs; this report could provide much of the necessary general information.

In addition, communication programs need tailor-made evaluation guidance, and many programs have insufficient budgets to hire experts. One solution would be to fund an expert advisory committee to consult on evaluation issues with communication programs that are small or have small budgets, but that have important public health implications.

Research topics that merit further study have been mentioned throughout this report. In addition, the experts advised:

- Moving toward more standardized criteria for quasi-experimental evaluation
- Advocating for non-RCTs as a standard, not a poor cousin of RCTs
- Extending the recommendations in the Belmont report
- Spending 10% of communication resources on evaluation
- Developing criteria for campaign effectiveness
- Analyzing the advantages and disadvantages of defining cost-effectiveness as cost per unit of change in an outcome measure vs. cost per person reached
- Conducting media tracking for high-priority messages at a national level rather than at the individual campaign level to conserve resources
- Using GIS to map access to campaign products and activities to illuminate campaign trends
- Updating the communication research agenda at CDC in each CIO
- Studying the work of Terrance Shimp, Erica Austin and Rich Lutz to clarify the theoretical underpinnings for applying the notion of branding to public health. The research base for this application should also be better established
- Studying the connection between interpersonal and mass communication

- Creating menus of best practices such as the guidebook for evaluating tobacco counter-marketing campaigns (Porter, Farrelly, Sly, & Murphy, in preparation)

Although advances in advertising, marketing, and private sector communication technologies are often proprietary, these advances should be tracked as closely as possible by health communication evaluators. For example, some health behaviors can be measured by sales (e.g., sales of cigarettes or low-fat milk), and detailed scanner data that can track sales across the country are underutilized in health communication. Scanner technology has been imported into a non-sales contexts by Valente, Foreman, Junge & Valhov (1998) who put bar codes on needles so their movement through IDU networks could be tracked. Ratings of television, radio, print and internet reach are another underutilized commercial resource (Webster, Phalen & Lichty, 2000; Porter, Farrelly, Sly & Murphy, in preparation). Media message evaluation can take advantage of variation in exposure by media market as measured by gross ratings points to understand the impact of advertising on program outcomes (Snyder, Fleming-Milici, Sun, Strizhakova & Slater, 2004).

Some thought about what makes a good partnership with the private sector is in order; collaboration and staff exchange with the private sector should be two-way. Commercial entities can learn from the experience health communicators have had with complex outcomes. In turn, health communication evaluators may be surprised to learn that Public Relations firms have found documentation of return-on-investment elusive. This exchange could be facilitated by joint best practices symposia and conferences with AAAA, the Association for Advertising Research, the Ad Council and AAPOR.

The parting recommendation the expert panel members made was to other evaluators who are in a position to present their findings to policy makers. Applied researchers were urged to acknowledge the limitations of their data. Scientific results are only one of several inputs into policy decisions. Characterizing the nature and degree of uncertainty in a set of findings helps an evaluator maintain credibility in the face of contradictory data from future studies, and helps the decision-maker give the data appropriate weight in a policy decision.

References

Ajzen, I. & Fishbein, M. (1980). *Understanding attitudes and predicting social behavior*. Englewood Cliffs, NJ: Prentice-Hall.

Allison, P.D. (2000). Multiple imputation for missing data: A cautionary tale. *Sociological Methods & Research, 28(3),* 301-309.

Anderson, E. (2000, Spring). Beyond Homo economicus: New developments in theories of social norms. *Philosophy and Public Affairs*, 29 (2), 170-200.

Andreasen, A.R. (1985) 'Backward' marketing research. *Harvard Business Review, May-June*, 176-182.

Babrow, A.S. (2001, Sept.). Uncertainty, value, communication, and problematic integration. *Journal of Communication*, 51 (3),: 553-573.

Bandura A. (1986) *Social Foundations of Thought and Action.* Englewood Cliffs, N.J.: Prentice-Hall.

Blum, H.L. (1976). Planning for health, development and application of social science change theory. *International Journal of Epidemiology*, 5 (2), 209-210.

Blumer, J.G. (1979). Role of Thoery in Uses and Gratifications Studies. *Communication Research*, 6 (1), 9-36.

Bollen, K.A., D.K. Guilkey, T.A. Mroz (1995). Binary outcomes and endogenous explanatory variables: tests and solutions with an application to the demand for contraceptive use in Tunisia. *Demography*, 32(1), 111-131.

Boulay, M., Storey, J.D., & Sood, S. (2002) Indirect exposure to a family planning mass media campaign in Nepal. *Journal of Health Communication, 7(5)*, 379-399.

Brehm, J.W. (1966). *A theory of psychological reactance*. New York: Holt, Rinehart, & Winston.

Bronfrenbrenner, U. (1979). *The ecology of human development*. Cambridge, MA: Harvard University Press.

Carli, L.L. & Eagly, A.H. (2001, Winter). Gender, hierarchy, and leadership: An introduction . *Journal of Social Issues*, 57 (4), 629-636.

Chen, H.T. (1990). *Theory-Driven Evaluation*, Sage.

Cole, G.E., (1999). Advancing the Development and Application of Theory-Based Evaluation in the Practice of Public Health. *American Journal of Evaluation*, Vol 20, No 3, pp. 453-470.

David, B. & Turner, J.C. (1996). Studies in self-categorization and minority conversion: Is being a member of the out-group an advantage? *British Journal of Social Psychology, 35*, Part 1: 79-199.

Des Jarlais, D.C., Lyles, C., Crepaz, N., and the Trend Group (2004). Improving the reporting quality of nonrandomized evaluations of behavioral and public health interventions: The TREND statement. *American Journal of Public Health, 94*, 361-366.

Dearing, J.W. (2004). Improving the state of health programming by using diffusion theory. *Journal of Health Communication,* 9, 21-36 Suppl. 1.

DiClemente, R.J., Crosby, R.A., Kegler, M.C. (eds) (2002). *Emerging Theories in Health Promotion Practice and Research*. San Francisco: Jossey-Bass.

Eveland, W.P. (1997). Interactions and nonlinearity in mass communication: Connecting theory and methodology. *Journalism & Mass Communication Quarterly*, 74 (2), 400-416.

Fazio, R.H. (1986). Attitude accessibility as a moderator of the attitude-perception and attitude-behavior relations: An investigation of the 1984 presidential election. *Journal of Personality and Social Psychology, 51*, 505-514.

Fazio, R.H. (1989). On the power and functionality of attitudes: The role of attitude accessibility. In A.R. Pratkanis, S.J. Breckler, & A.G. Greenwald (Eds.), *Attitude structure and function* (pp. 153-179). Hillsdale, NJ: Erlbaum.

Figueroa, M. E., Bertrand, J.T. & Kincaid, D. L. (2002). Evaluating the impact of communication programs: Summary of an expert meeting organized by the measure evaluation project and the population communication services project. *Belmont Conference Center*, Elkridge, MD, prepared for USAID.

Foley, D. & Pechmann, C. (2004). The national youth anti-drug media campaign copy test system. *Social Marketing Quarterly, 10(2),* 34-42.

Fortmann, S.P., Taylor, C.B., Flora, J.A. & Jatulis, D.E. (1993). Changes in adult cigarette smoking prevalence after 5 years of community health education: the Stanford Five-City Project. *American Journal of Epidemiology, 137(1),* 82-96.

Foster-Cohen, S.H.. (2004, July). Relevance Theory, Action Theory and second language communication strategies. *Second Language Research, 20(3),* 289-302.

Fridinger, F. & Kirby, S. (2002) The "Doublemint" factor: Issues and challenges in marketing nutrition and physical activity behaviors in 1 program. Social Marketing Quarterly, 8(4), 1-13.

Gerbner, G. & Gross, L. (1976). Living with television: The violence profile. *Journal of Communication, 26,* 76.

Glantz, K., Lewis, FM. & Rimer, B. (1997). *Health behavior and health education: theory, research, and practice,* 2nd edition. San Francisco: Jossey-Bass Publishers.

Goodman, R.M., McLeroy, K.R., Steckler, A.B., & Hoyle, R.H. (1993). Development of level of institutionalization scales for health promotion programs. *Health Education Quarterly, 20(2),* 161-178.

Green, L.W. & Kreuter, M.W. (1991). *Health Promotion Planning: An Educational and Environmental Approach* (Second Ed.). Mountain View, Cal.: Mayfield.

Grunig, J.E. & Hunt, T. (1984). *Managing public relations*. New York: Holt, Rinehart and Winston.

Grunig, L.A. & Turner, J.H. (1989). A Theory of social interaction. *Public Relations Review*, 15 (2), 71-72.

Hennekens, C.H. & Buring, J.E. (1987) *Epidemiology in Medicine*. Boston: Little Brown.

Himmelstrand, U. (1981). Innovative processes in social change theory: method and social practice. *International Social Science Journal*, 33 (2), 227-247.

Hedeker, D. & Gibbons, R.D. (1997). Application of random-effects pattern-mixture models for missing data in longitudinal studies. *Psychological Methods, 2(1),* 64-78.

Hornik, R.C. (2001). Epilogue: Evaluation design for public health communication programs. In R.C. Hornick (Ed.) *Public Health Communication: Evidence for Behavior Change*, Mahwah, NJ: Lawrence Erlbaum Associates.

House, R.J. &, Aditya, R.N. (1997). The social scientific study of leadership: Quo vadis? *Journal of Management*, 23 (3), 409-473.

Institute of Medicine of the National Academies (IOM). (July 8, 2002). Speaking of Health: Assessing Health Communication Strategies for Diverse Populations. *Board on Neuroscience and Behavioral Health.*

Kawachi, I. & Berkman, L.F. (2001, Sept). Social ties and mental health. *Journal of Urban Health-Bulletin of the New York Academy of Medicine*, 78 (3), 458-467.

Kincaid, D.L. (2004). From innovation to social norm: Bounded normative influence. *Journal of Health Communication*, 9: 37-57, Suppl. 1.

Kincaid, D.L. (2000a). Social networks, ideation and contraceptive behavior in Bangladesh: A longitudinal . *Social Science and Medicine, 50(2),* 215-231.

Kincaid, D.L. (2000b). Mass media, ideation and behavior: A longitudinal analysis of contraceptive change in the Philippines. *Communication Research, 27,* 764-763.

Kondo, A. (1975). Planning for health, development and application of social change theory. *American Journal of Public Health,* 65 (1), 86-87, 1975.

Little, R. J. A. & Rubin, D. A. (1987). *Statistical analysis with missing data*. New York: John Wiley & Sons.

Martin, P.Y. (2004, June). Gender as social institution. *Social Forces*, 82 (4), 1249-1273.

McGuire, W. (1981). Theoretical foundations of campaigns. In Ronald Rice and William Paisley (eds.), *Public Communication Campaigns*, Thousand Oaks, CA: Sage.

Murray, D.M. (1998). Design and analysis of group-randomized trials. New York: Oxford University Press.

Murphy, P. (1996). Chaos theory as a model for managing issues and crises. *Public Relations Review*, 22, 95-114.

NCI (2003). *Theories at a glance*. Available on the internet at: http://www.cancer.gov/cancerinformation/theory-at-a-glance

Palmgreen, P., Donohew, L., Lorch, E.P., Hoyle, R.H., & Stephenson, M.T. (2002) Television campaigns and sensation seeking targeting of adolescent marijuana use: A controlled time series approach, In R.C. Hornik (ed.) Public health communication: Evidence for behavior change (pp. 35-56). Mahway, NJ: Larwrence Erlbaum Associates.

Palmgreen., P., Donohew, L., Lorch, E., Hoyle, R., & Stephenson, M. (2001). Television campaigns and adolescent marijuana use: Tests of sensation seeking targeting. *American Journal of Public Health*, 91, 292-295.

Petty, R.E., & Cacioppo, J.T. (1979). Elaboration likelihood model: Issue involvement can increase or decrease persuasion by enhancing message-relevant cognitive responses. *Journal of Personality and Social Psychology, 37,* 1915-1926.

Plewczynski, D. (1998). Landau theory of social clustering. *Physica HYSICA A, 261 (3-4)*, 608-617.

Porter, S., Farrelly, M., Sly, D., & Murphy, R. (in preparation). Guidebook for evaluating tobacco counter-marketing campaigns.

Prochaska, J.O., DiClemente, C.C., Norcross, J.C. (1992). In search of how people change: *Applications to addictive behaviors. American Psychologist, 47,* 1102-1114.

Rice, R.E., & Paisley, W.J. (Eds.) (1981). Public Communication Campaigns. *Beverly Hills*: Sage.

Ridgeway, C.L. & Correll, S.J. (2004, August). Unpacking the gender system: A theoretical perspective on gender beliefs and social relations . *Gender & Society*, 18 (4), 510-531.

Rogers, E.M. & Kincaid, D.L. (1981) Communication Networks: Toward a New Paradigm for Research. New York: Free Press.

Rogers, E.M. (1995). *Diffusion of innovations* , New York: Free Press.

Rosenbaum, P.R. & D.B. Rubin. (1983) Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *The Journal of the Royal Statistical Society,* Series B, 45, 212-8.

Roth, P. (1994). Missing data: A conceptual review for applied psychologist. *Personnel Psychology, 47,* 537-560.

Rothman, P. & Salovey, P. (1997). Shaping perceptions to motivate healthy behavior: The role of message framing. *Psychological Bulletin, 121 (1),* 3-19.

Shadish, W., Cook, T. & Leviton, L. (1990). *Foundations of Program Evaluation: Theories of Practice,* Sage.

Shimp, T.A. (1981) Attitude toward the ad as a mediator of consumer brand choice. *Journal of Advertising, 10(2),* 9-15.

Slater, M.D. (2003, March). Alienation, aggression, and sensation seeking as predictors of adolescent use of violent film, computer, and website content. *Journal of Communication*, 53 (1), 105-121.

Slater, M.D. & Kelly, K.J. (2002). Testing alternative explanations for exposure effects in media campaigns: The case of a community-based, in-school media drug prevention project. *Communication Research, 29,* 367-389.

Smedley, B.D., Stith, A.Y., & Nelson, A.R. (2003) Unequal treatment: Confronting racial and ethnic disparities in health care. NAS Institute of Medicine: Washington, DC.

Snyder, L.B. & Hamilton, M.A. (1999). When evaluation design affects results: Meta-analysis of evaluations of mediated health campaigns. Presented at the annual conference of the Association for Education in Journal and Mass Communication, New Orleans, August.

Snyder, L.B., Hamilton, M.A., Mitchell, E.W., Kiwanuka-Tondo, J., Fleming-Milici, F., & Proctor, D. (2004). A meta-analysis of the effect of mediated health communication campaigns on behavior change in the United States. *Journal of Health Communication, 9/Supplement 1*,71-96.

Southwell, B.G., Barmada, C.H., Hornik, R.C., & Maklan, D.M. (2002) Can we measure encoded exposure? Validation evidence from a national campaign. *Journal of Health Communication, 7(5),* 445-453.

Storey, J.D., M. Boulay, Y. Karki, K. Heckert, & D.M. Karmacharya. (1999). Impact of the integrated radio communication project in Nepal, 1994-1997. *Journal of Health Communication, 4(4),* 271-94).

Stryker, J.E. (2003). Media and marijuana: A longitudinal analysis of news media effects on adolescents' marijuana use and related outcomes, 1977-1999. *Journal of Health Communication, 8(4),*305-28.

Stufflebeam, Daniel (2001). *Evaluation Models: New Directions in Evaluation*, Jossey-Bass.

Sun, Y.F. (2003) Translating cultural differences. *Perspectives- Studies in Translatology, 11(1),* 25-36.

Terry, D.J., Hogg, M.A, & White, K.M.. (1999). The theory of planned behaviour: Self-identity, social identity and group norms. *British Journal of Social Psychology, 38,* 225-244, Part 3.

Thapa, B., A.R. Graefe & J.D. Absher. (2002). Information needs and search behaviors: A comparative study of ethnic groups in the Angeles and San Bernardino National Forests, California. *Leisure Sciences, 24 (1),* 89-107.

Thurman, P.J., Plested, B.A., Edwards, R.W., Foley, R. & Burnside, M. (2003) Community readiness: The journey to community healing. *Journal of Psychoactive Drugs, 35(1),* 27-31.

UNAIDS (2000). National AIDS Programmes: A guide to monitoring and evaluation (go to www.unaids.org to find an example of the WHO indicator approach), click on resources, then publications, and search on monitoring and evaluation.

Valente, T.W. (1995) *Network Models of the Diffusion of Innovations*. Cresskill, NJ: Hampton Press.

Valente, T.W. (2002). *Evaluating Health Promotion Programs.* New York: Oxford University Press.

Valente, T.W., Foreman, R.K., Junge, B., & Valhov, D. (1998). Satellite exchange in the Baltimore needle exchange program. Public Health Reports, 113(S1): 91-96.

Von Bertalanffy, L. (1967). General systems theory. (In) N.J. Demerath, III & R.A. Peterson (Ed.), *System, change, and conflict* (pp. 115-129). New York: The Free Press.

Webster, J.G., Phalen, P.F., & Lichty, L.W. (2000) *Ratings analysis*, 2nd Edition. Mahwah, NJ: Lawrence Erlbaum.

Wein, N.D. (1987). Unrealistic optimism about susceptibility to health problems: conclusions from a community-wide sample. *Journal of Behavioral Medicine*, *10*, 481-500.

Williams-Piehota, P., Schneider, T.R., Pizarro, J, Mowad, L., & Salovey, P. (2003). Matching health messages to information-processing styles: Need for cognition and mammography utilization. *Health Communication, 15 (4),* 375-392.

Willis, G.B. (1999). Cognitive interviewing: A "how to" guide. Presented at the annual meeting of the American Statistical Association by R.A. Caspar, J.T. Lessler, and G.B.Willis. Available on the internet at http://appliedresearch.cancer.gov/areas/cognitive/guides.html

Witte, K. (1992). Putting the fear back into fear appeals: The extended parallel process model. *Communication Monographs, 59,* 330-349.

Wothke, W. (1998). Longitudinal and multi-group modeling with missing data. In T.D. Little, K.U. Schnabel, & J. Baumert (Eds.) *Modeling longitudinal and multiple group data: Practical issues, applied approaches and specific examples.* Mahwah, NJ; Lawrence Erlbaum Associates.

Yin, R. (2003). *Case study research: design and methods*. Thousand Oaks, California: Sage Publications.

Yanovitzky, I. (2002) Effect of news coverage on the prevalence of drunk-driving behavior: evidence from a longitudinal study. *Journal of Studies on Alcohol, 63(3),* 342-351.

Zaza, S., Wright-de Aguero, L., Briss, P.A., Turman, B.I., Hopkins, D.P., Hennessy, M.H., Sosin, D. M., Anderson, L., Carande-Kulis, V.G., Teutsch, S.M.,

Papaioanou, M., (2000) Task Force on Community Preventive Services. Data collection instrument and procedure for systematic reviews in the *Guide to Community Preventive Services*. *American Journal of Preventive Medicine, 18(1S),* 44-74.

Appendix A: Participant List

External Evaluation Experts:

Baur, Cynthia, Ph.D.
Health Communication and e-Health Advisor
Office of Disease Prevention and Health Promotion
U.S. Department of Health and Human Services

Bernhardt, Jay, Ph.D.
Assistant Professor
Behavioral Science and Health Education
Rollins School of Public Health
Emory University

Cho, Hyunyi, Ph.D.
Assistant Professor
Department of Communication
Purdue University

Cotton, David, Ph.D.
ORC/Macro International

Denniston, Robert, Ph.D.
White House Office of National Drug Control Policy

Farrelly, Matthew, Ph.D.
Director, Public Health Economics and Policy Research Program
RTI International

Figueroa, Maria Elena, Ph.D.
Chief, Research and Evaluation Division
Johns Hopkins University, Center for Communication Programs

Hornik, Robert, Ph.D.
Professor of Communication and Health Policy
Director, Center of Excellence in Cancer Communication Research
Anneberg School for Communication
University of Pennsylvania

Kreps, Gary, Ph.D.
Eileen and Steve Mandell Endowed Chair in Health Communication
Professor and Chair, Department of Communication
George Mason University

Middlestadt, Susan, Ph.D.
Associate Professor
Department of Applied Health Science
Indiana University

Parrott, Roxanne, Ph.D.
Professor of Communication Arts & Sciences
Pennsylvania State University

Slater, Michael, Ph.D.
Professor
Department of Journalism and Technical Communication
Colorado State University

Snyder, Leslie, Ph.D.
Associate Professor
Communication Sciences
University of Connecticut

Storey, Douglas, Ph.D.
Associate Director for Program Research
The Health Communication Partnership
Johns Hopkins Bloomberg School of Public Health
Center for Communication Programs

CDC meeting staff:

Abbatangelo, Jodie, M.A., Sc.M.
ORISE Fellow
Harvard University

Chapel, Tom. M.S.
Office of Strategic Initiatives

Cole, Galen, Ph.D.
Center for Health Marketing,
Division of Health Communication

Jones, Vivian
Center for Health Marketing,
Division of Health Communication

Kennedy, May, Ph.D.
Center for Health Marketing,
Division of Health Communication

Prue, Christine, Ph.D.
National Center and Division of Birth Defects and Developmental Disabilities

Polonec, Lindsey, M.A.
ORISE Fellow

CDC Participants:

Anderton, John, Ph.D.
National Center for HIV, STD, & TB Prevention
Office of Media Relations

Forsythe, Ann, Ph.D.
National Center for Chronic Disease Prevention and Health Promotion,
Division of Nutrition and Physical Activity

Galavotti,Christine, Ph.D.
National Center for Chronic Disease Prevention and Health Promotion/
Division of Reproductive Health & Global AIDS Program

Guenther-Grey, Carolyn, M.S.
National Center for HIV, STD, and TB Prevention
Prevention Research Branch

Huhman, Marion, Ph.D.
National Center for Chronic Disease Prevention and Health Promotion
VERB Campaign

Johnson, Wayne D., M.S.
National Center for HIV, STD, and TB Prevention,
Prevention Research Branch

Jorgenson, Cynthia, Ph.D.
National Center for Chronic Disease Prevention & Health Promotion,
Division of Cancer Prevention

Lackey, Cheryl, M.A.
Center for Health Marketing,
Division of Creative Services

Lewis, Sonya, M.A.
CDC/ORISE Fellow

McDivitt, Jude, Ph.D.
National Center for Chronic Disease Prevention & Health Promotion,
Division of Nutrition and Physical Activity

Nowak, Glen, Ph.D.
Division of Media Relations

Parvanta, Claudia, Ph.D.
Center for Health Marketing,
Division of Health Communication

Pollard, Bill, Ph.D.
Center for Health Marketing,
Division of Health Communication

Robinson, Susan, M.S.
ATSDR
Office of Media Relations

Vanderford, Marsha, Ph.D.
Center for Health Marketing,
Division of Health Communication

Appendix B: Pre-Panel Reading List

Chapel, T. J. (2000). Introduction to program evaluation: Participant guide and case studies. Drawn from Veney, J.E. & Kaluzny, A.D. (1991) *Evaluation and decision-making for health services*, 2nd Ed. Ann Arbor, MI: Health Administration Press.

Figueroa, M. E., Bertrand, J.T. & Kincaid, D. L. (2002). *Evaluating the impact of communication programs: Summary of an expert meeting organized by the MEASURE Evaluation Project and the Population Communication Services Project.* Belmont Conference Center, Elkridge, MD, prepared for USAID.

Freimuth, V., Cole, G. & Kirby, S. D. Issues in evaluating mass-media health communication campaigns. In: I. Rootman, M. Goodstadt, B. Hyndman, D.V. McQueen, L. Potvin, J. Springett, & E. Ziglio (eds.) *Evaluation in health promotion: Principles and perspectives.* Copenhagen: WHO Regional Publications, European Series, No. 92

Hornik, R.C. (2001). Epilogue: Evaluation design for public health communication programs. In R.C. Hornick (Ed.) *Public Health Communication: Evidence for Behavior Change*, Mahwah, NJ: Lawrence Erlbaum Associates.

Koplan, J. P. (1999). Framework for program evaluation in public health. *Morbidity and Mortality Weekly Report, 48,* 1-40.

UNAIDS (2000) National AIDS programmes: A guide to monitoring and evaluation. Document #9. (retrieved May 3rd, 2004 at www.unaids.org by clicking on resources, then publications, then searching on monitoring and evaluation.

US DHHS (July, 2003) *Communicating health: Priorities and strategies for progress.* Objective 11.3: Research and Evaluation of Health Communication Programs, 61-73.

Appendix C: "Belmont 8" Criteria for Claiming Impact

1. Observation of a change in the expected outcome
2. Correlation between that change and exposure to the program
3. Evidence that exposure occurred before the observed change (time-order)
4. No evidence of confounding variables that may have accounted for the change
5. Observation of a large, abrupt impact (magnitude): if promotion, easy and immediate/ abrupt; if behavior that needs longer time-period, intermediate outcomes; maturity of program for some behaviors
6. Evidence of a causal connection (proximity and theoretical coherence)
7. Evidence that impact increases in proportion to level/duration of exposure (dose response)
8. Consistency with previous program research (replication with variation)

Sources: Piotrow et al., 1997; Bertrand & Kincaid, 1996; Hill, 1971.